AWS Solutions Architecture Notes

Your Name

Date:

Contents

Chapter 1

Course Introduction

Pre-Introduction The purpose of these notes are two-fold.

- Prepare for this exam in the future.
- Gain a high level overview of the services offered by one of the main cloud providers.
- Additionally it could be useful for generating Terraform examples.

Chapter 2

Identity & Federation

2.1 IAM - What should you know by now

- Users: Long term credentials - Groups - Roles: short-term credentials, uses STS - EC2 Instance Roles: Uses the EC2 metadata service. One role of a time per instance - Service Roles: API Gateways, CodeDeploy etc - Cross Account roles - Policies - AWS Managed - Customer Managed - Inline Policies - Resource Based Policies (S3 Bucket, SQS queues, etc......)

2.1.1 IAM Policies Deep Dive

- Anatomy of a policy: JSON doc with Effect, Action, Resource, Conditions, Policy Variables - Explicit DENY has precedence over ALLOW - Best practice: use least privilege for maximum security - Access Advisor: See permissions granted and when last accessed - Access Analyser: Analyse resources that are shared with external entity - Navigate Examples at:

2.1.2 IAM AWS Managed Policies - Administrator Access example

Listing 2.1: Sample JSON Data

```
{
"Version": "2012-10-17",
"Statement": [
{
"Effect": "Allow",
"Action": "*",
"Resource": "*"
}
]
}
```

2.1.3 IAM Policies Conditions

2.1.4 IAM Policies Variables and Tags

2.1.5 IAM Roles vs Resource Based Policies

- Attach a policy to a resource (example: S3 bucket policy) versus attaching of a using a role as a proxy - When you assume a role (user, application or service), you give up your original permissions and take the permissions assigned to the role. - When using a resource-based policy, the principal doesn't have to give up any permissions - Example: User in account A needs to scan a DynamoDB table in Account A and dump it in an S3 bucket in Account B

2.1.6 IAM Permissions Boundaries

- IAM permission boundaries are supported for users and roles (not groups) - Advanced feature to use a managed policy to set the maximum permissions an IAM entity can get - Can be used in combinations of AWS organisations SCP

2.1.7 Use cases

- Delegate responsibilities to non administrators within their permission boundaries, for example create create new IAM users
- Allow developers to self-assign policies and managed their own permissions while making sure they can't 'escalate' their privileges (i.e make themselves admin)
- Useful to restrict one specific user (instead of a whole account using Organisations and SCP)

2.1.8 IAM Access Analyser

- Find out which resources are shared externally - S3 Buckets - IAM Roles - KMS Keys - Lambda Functions and Layers - SQS queues - Secrets Manager Secrets - Define Zone ofTrust = AWS Account or AWS Organisation - Access outside of zone trusts =¿ findings

2.1.9 IAM Access Analyser

- IAM Access Analyser Policy Validation - Validates your policy against IAM policy grammar and best practices - General warnings, security warnings errors suggestions - Provides actionable recommendations - IAM Access Analyser Policy Generation - Generates IAM policy based on access activity - CloudTrail logs is reviewed to generate the policy with the fine-grained permissions and the appropriate Actions and Services - Reviews CloudTrail logs for up to 90 days

2.2 STS

2.2.1 Using STS to Assume a Role

- Define an IAM Role within your account or cross-account - Define which principals can access this IAM role - Using AWS STS (Security Token Service) to retrieve credentials and impersonate the IAM role you have access to (AssumeRole API) - Temporary credentials can be valid between 15 to 12 hours.

2.2.2 Assuming a Role with STS

- Provide access for an IAM user in one AWS account that you own to access resources in another account that you own. - Provide access to IAM users in AWS accounts owned by third parties - Provide access for services offered by AWS to AWS resources - Provide access for externally authenticated users (identity federation) - Ability to revoke active sessions and credentials for a role (by adding a policy using a time statement - AWSRevokeOlderSessions)

When you assume a role (user, application or service), you give up your original permissions and take the permissions assigned to the role.

2.2.3 Providing Access to an IAM User in Yours or Another AWS Account That You Own

- You can grant your IAM users permissions to switch to roles within your AWS account or to roles defined in other AWS accounts that you own. - Benefits: - You must explicitly grant your users permission to assume the role - Your users must actively switch to the role using the AWS management console or assume the role using the AWS CLI or AWS API - You can add multi-factor authentication (MFA) protection to the role so that only users who sign in with an MFA device can assume the role. - Least Privilege + auditing using CloudTrail

2.2.4 Providing Access to AWS Accounts Owned by Third Parties

- Zone of trust = accounts, organisations that you own - Outside Zone ofTrust = 3rd parties - Use IAM Access Analyser to find out which resources are exposed - For granting access to a thrid party: - The third part AWS account ID - An External ID (secret between you and the third party) - To uniquely associate with the role between you and 3rd party - Must be provided when defining the trust and when assuming the role - Must be chosen by the third party - Define permissions in the IAM policy

Session Tags in STS

- Tags that you pass when you assume an IAM Role or federate user in STS - aws:PrincipalTag Condition - Compares the tags attached to the principal making the request with the tag you specified in the policy. - Example: Allow a principal to pass session tags only if the principal making the reqeust has the specified tags.

STS Important APIs

- AssumeRole: access a role within your account or cross-account - AssumeRoleWithSAML: return credentials for users logged with SAML - AssumeRoleWithWebIdentity: return creds for users logged with an IdP - Example providers include Amazon Cognito, Login with Amazon, Facebook, Google or any OpenID Connect-compatible identity provider. - AWS Recommends using Cognito - GetSessionToken: for MFA from a user or AWS account root user - GetFederationToken: obtain temporary creds for a federated user, usually a proxy app that will give the creds for a federated user, usually a proxy app that will give the creds to a distributed app inside a corporate network.

2.2.5 Identity Federation in AWS

- Give users outside of AWS permissions to access AWS resources in your account - You don't need to create IAM Users (user management is outside AWS) - Use cases: - A corporate has its own identity system (e.g Active Directory) - Web / Mobile application that needs access to AWS resources - Identity Federation can have many flavours: - SAML 2.0 - Custom Identity Broker - Web Identity Federation With(out) Amazon Cognito - IAM Identity Centre

SAML 2.0 Federation

- Security Assertion Markup Language 2.0 (SAML 2.0) - Open standard used by many identity providers (e.g ADFS) - Supports integration with Microsoft Active Directory Federations Services (ADFS) - Or any SAML 2.0 - compatible IdPs with AWS - Access the AWS Console, AWS CLI or AWS API using temporary credentials - No need to create IAM Users for each of your employees - Need to setup a trust between AWS IAM and SAML 2.0 Identity Provider (both ways) - Under-the-hood: Uses the STS API AssumeRoleWithSAML - SAML 2.0 Federation is the ”old way”, IAM Identity Center Federation is the new managed and simpler way

SAML 2.0 Federation - AWS API Access

SAML 2.0 Federation - AWS Console Access

SAML 2.0 Federation - Active Directory DS (ADFS)

Custom Identity Broker Application

- Use only if identity provider is NOT compatible with SAML 2.0 - The identity broker is NOT compatible with SAML 2.0 - The identity broker must determine the appropriate IAM Role - Uses the STS API AssumeRole or GetFederationToken

Web Identity Federation - Without Cognito

- Not recommended by AWS - use cognito instead

Web Identity Federation - With Cognito

- Preferred over for Web Identity Federation - Create IAM Roles using Cognito with the least privilege needed - Build trust between the OIDC IdP and AWS - Cognito benefits: - Supports anonymous users - Supports MFA - Data Synchronisation - Cognito replaces a Token Vending Machine (TVM)

Web Identity Federation - IAM Policy

- After being authenticated with Web Identity Federation, you can identify the user with an IAM policy variable - Example:

2.2.6 AWS Directory Services (AD)?

- Found on any Windows Server with AD Domain Services - Database of objects: User Accounts, Computers, Printers, File Shares, Security Groups - Centralised security management, create account, assign permissions - Objects are organised in trees - A group of trees

What is ADFS (AS Federation Services)?

- ADFS provides Single Sign-On across applications - SAML across 3rd party: AWS Console, Dropbox, Officer365, etc.....

AWS Directory Services

- AWS Managed Microsoft AD - Create your own AD in AWS, manage users, locally supports MFA - Establish ”trust” connection with your own on premises - AD Connector - Directory Gateway (proxy) to redirect to on-premises AD, supports MFA - Users are managed on the on-premise AD - Simple AD - AD-compatible managed directory on AWS - Cannot be joined with on-premise AD

AWS Directory Services and AWS Managed Microsoft AD

- Managed Service: Microsoft AD in your AWS VPC - EC2 Windows Instances: - EC2 Windows instances can join the domain and run traditional AD applications (sharepoint, etc) - Seamlessly Domain Join Amazon EC2 Instances from Multiple Accounts and VPCs - Integrations - RDS for SQL Server, AWS Workspaces, Quicksight....... - AWS SSO to provide access to third party applications - Standalone repository in AWS or jointed to on-premises AD - Multi AZ deployment of AD in 2 AZ, of DC (Domain Controllers) can be increased for scaling - Automated backups - Automated Multi-Region replication of your directory

AWS Microsoft Managed AD - Integrations

Connect to on-premise AD

- Ability to connect your on-premise Active Directory to AWS Managed Microsoft AD - Must establish a Direct Connection (DX) or VPN connection - Can setup three kinds of forest trust - One-way trust: AWS -¿ On-Premise - One-way trust: On-Premise -¿ AWS - Two-way forest: trust - AWS -¿ On-Premise - Forest trust is different than synchronisation (replication is not supported)

Solution Achitecture: Active Directory Replication

- You may want to create a replica of your AD on EC2 in the cloud to minimise latency of in case DX or VPN goes down - Establish trust between the AWS Manged Microsoft AD and EC2

AWS Directory Services AD Connector

- AD Connector is a directory gateway to redirect director requests to your on premises Microsoft Active Directory - No caching capability - Managed users solely on-premise, no possibility of setting up a trust - VPN or Direct Connect - Doesn’t work with SQL Server, doesn’t do seamless joining, can’t share director.

AWS Directory Services Simple AD

- Simple AD is an inexpensive Active Directory-compatible service with the common directory features - Supports joining EC2 instances, manage users and groups - Does not support MFA, RDS SQL server, AWS SSO - Small: 500 users, large: 5000 users - Powered by Samba 4, compatible with Microsoft AD - lower cost, low scale, basic AD compatible or LDAP compatibility - No trust relationship

2.2.7 AWS Orgnisations

AWS Organisations - OrgnizationAccountAccessRole

- IAM role which grants full administrator permissions in the Member account to the Management account - Used to perform admin tasks in the Member accounts (e.g - creating IAM users) - Could be assumed by IAM users in the Management account - Automatically added to all new Member account created with AWS organisations - Must be created manually if you invite an existing Member account

Multi Account Strategies

- Create accounts per department, per cost centre, per dev / test / prod, based on regulatory restrictions (using SCP), for better resource isolation (ex:VPC), to have separate per-account service limits isolated account for logging. - Multi Account vs. On Account MultiVPC - Use tagging standards for billing purposes - Enable CloudTrail on all accounts, send logs to central S3 account - Send CloudWatch logs to central logging account - Strategy to create an account for security

Orgnisational Units (OU) - Examples

AWS Organisation - Feature Modes

- Consolidated billing features: - Consolidated Billing across all accounts - single payment method - Pricing benefits from aggregated usage (volume discount for EC2, S3.....)

All Features (Default)

- Includes consolidated billing features, SCP - Invited accounts must approve enabling all features - Ability to apply an SCP to prevent member accounts from leaving the org - Can't switch back to Consolidated Billing Features only

AWS Organisations - Reserved Instances

- For billing purposes, the consolidated billing features of AWS organisations treats all the accounts in the organisation as one account. - The means that all accounts in the organisation can receive the hourly cost benefit of Reserved Instances that are purchased by any other account. - The payer account (Management account) of an organisation can turn off Reserved Instance (RI) discount and Saving Plans discount sharing for any accounts in that organisation, including the payer account. - This means that RIs and Saving Plans discounts aren't shared between any accounts that have sharing turned off. - To share an RI or Savings Plans discount with an account, both accounts must have sharing turned on.

AWS Organisation - Moving Accounts

- Remove the member account from the AWS organisation - Send an invite to the member account from the AWS organisation - Accept the invite to the new Organisation from the member account.

Service Control Policies (SCP)

- Define allowlist or blocklist IAM actions - Applied at the OU or Account level - Does not apply to the Management Account - SCP is applied to all the Users and Roles in the account, including Root user - The SCP does not affect Service-linked roles - Service-linked roles enable other AWS services to integrate with AWS organisations and can't be restricted by SCP's - SCP must have an explicit Allow from the root of each OU in the direct path to the target account (does not allow anything by default) - Use cases: - Restrict access to certain services (for example: can't use EMR) - Enforce PCI compliance by explicitly disabling services.

SCP Hierarchy

- Management Account - Can do anything (no SCP apply) - Account A - Can do anything - EXCEPT S3 (explicit Deny from Sandbox OU) - EXCEPT EC2 (explicit deny) - Account B and C - Can do anything - EXCEPT S3 (explicit Deny from Sandbox OU) - Account D - Can access EC2 - Prod OU and Account E and F

SCP Examples - Blocklist and Allowlist strategies

IAM Policy Evaluation Logic

Restricting Tags with IAM policies

- You can restrict specific tags on AWS resources - Using the aws:TagKeys Condition Key - Validate the Tag Keys attached to a resource against the Tag Keys in the IAM Policy - Example: Allow IAM users to create EBS Volumes only if it has the ”Env” and ”CostCenter” Tags - Use either ForAllValues (must have all keys) or ForAnyValue (must have any of these keys at a minimum)

Using SCP to restrict creating resources without appropriate tags

- Prevent IAM Users/Roles in the affected member accounts from creating resources if they don't have a specific Tag

AWS Organisations - Tag Policies

- Helps you standardise tags across resources in an AWS organisation - Ensure consistent tags, audit tagged resources, maintain proper resources categorisation - You define Tag keys and their allowed values - Helps with AWS Cost Allocation Tags and Attribute-based Access Control - Prevent any non-compliant tagging operations on specified services and resources - Generate a report that lists all tagged/ non compliant resources - Use Amazon EventBridge to monitor non-compliant tags

AWS Organisation - AI Service Opt-out Policies

- Certain AWS AI services may use your content for continuous improvement of Amazon AI/ML services - Example: Amazon Lex, Amazon Comprehend, Amazon Polly......... - You can opt-out of having your content stored or used by AWS AI services - Create an Opt-out Policy that enforces this setting across all Member accounts and AWS Regions - You can opt-out all AI services or selected services - Can be attached to Organisation Root, specific OU or individual Member account

AWS Organisations - Backup policies

- AWS Backup enables you to create Backup Plans that define how to backup your AWS resources - JSON Documents that define backup plans across an AWS Organisation - Gives you granular control over backup up your resources (e.g backup frequency, time window, backup region,.....) - Can be attached to Organisation Root, specific OU or individual Member account - Immutable backup plans appear in Member accounts (view only)

Using SCP to Deny a Region aws:ReqeustRegion

2.2.8 AWS IAM Identity Center

AWS IAM Identity Center -successor to AWS Single Sign-On-

- One login (single sign-on) for all your - AWS accounts in AWS organisations - Business cloud applications (e.g, Salesforce, Box, Microsoft 365) - SAML2.0 enabled applications - EC2 windows instances

- Identity Providers - Built-in identity store in IAM identity center - 3rd party Active Directory (AD), OneLogin, Okta

- AWS IAM Identity Center - Login Flow

AWS IAM Identity Center

AWS IAM Identity Center - Fine-grained Permissions and Assignments

- Multi-Account Permissions - Manage access across AWS accounts in your AWS Organisation - Permission Sets - a collection of one or more IAM policies assigned to users and groups to define AWS access - Application Assignments - SSO access to many SAML 2.0 business applications (Salesforce, Box, Microsoft 365) - Provide required URLs, certificates and metadata - Attribute-Based Access Control (ABAC) - Fine-grained permissions based on users' attributes stored in IAM identity center identity store - Example: Cost center, title, locale - Use case: Define permission once, then modify AWS access by changing the attributes

AWS Control Tower

- Easy way to setup and govern a secure and compliant multi-account AWS environment based on best practices - Benefits: - Automate the set up of your environment in a few clicks - Automate ongoing policy management using guardrails - Detect policy violations and remediate them

- Monitor compliance through an interactive dashboard - AWS ControlTower runs on top AWS Organisations: - It automatically sets up AWS Organisations to organise accounts and implement SCPs (Service Control Policies)

AWS Controller Tower - Account Factory

- Automates account provisioning and deployments - Enables you to create pre-approved baselines and configuration options for AWS accounts in your organisation (e.g VPC default configuration, subnets, region, ...) - Uses AWS service catalog to provision new AWS accounts

AWS Control Tower - Detect and Remediate Policy Violations

- Guardrail - Provide ongoing governance for your Control Tower environment (AWS Accounts) - Preventive - using SCPs (e.g Disallow Creation of Access Keys for the Root User) - Detective - users AWS Config (e.g Detect Whether MFA for the Root User is Enabled) - Example: identify non-compliant resources (e.g, untagged resources)

AWS Control Tower - Guardrails Levels

- Mandatory - Automatically enabled and enforced by AWS control tower - Example: Disallow public Read access to the Log Archive account - Strongly Recommended - Based on AWS best practices (optional) - Example: Enable encryption for EBS volumes attached to EC2 instances - Elective - Commonly used by enterprises (optional) - Examples: Disallow delete actions without MFA in S3 buckets

2.3 AWS Resource Access Manager (RAM)

- Share AWS resources that you own with other AWS accounts - Share with any account or within your Organisation - Avoid resource duplication! - VPC Subnets - Allow to have all the resources launched in the same subnets - Must be from the same AWS organisations - Cannot share security groups and defaultVPC - Participants can manage their own resources in there - Participants can’t view, modify, delete resources that belong to other participants or the owner - AWS Transit Gateway - Route 53 (Resolver Rules, DNS Firewall Rule Groups) - License Manager Configurations

AWS Resource Access Manager (RAM)

- Aurora DB Clusters - ACM Private Cerficiate Authority - CodeBuild Project - EC2 (Dedicated Hosts, Capacity Reservation) - AWS Glue (Catalog, Database, Table) - AWS Network Firewall Policies - AWS Resources groups - Systems Manager Incident Manager (Contacts, Response Plans) - AWS Outposts (Outpost, Site)

Resource Access Manager - VPC example

- Each account...... - Is responsible for it own resources - Cannot view modify or delete other resources in other accounts - Network is shared so..... - Anything deployed in the VPC can talk to other resources in the VPC - Applications are accessed easily across accounts, using a private

IP - Security groups from other accounts can be referenced for maximum security - Use cases - Applications within the same trust boundaries - Applications with a high degree of interconnectivity

Resource Access Manager Managed Prefix List

- A set of one or more CIDR blocks - Makes it easier to configure and maintain Security Groups and Route Tables - Customer-Managed Prefix List - Set of CIDRs that you define and manage by you - Can be shared with other AWS accounts or AWS Organisation - Modify to update many security groups at once - AWS-Managed Prefix List - Set of CIDRs for AWS services - You can’t create, modify, share or delete them.

Resource Access Manger Route 53 Outbound Resolver

- Helps you scale forwarding rules to your DNS in case you have multiple accounts and VPC

2.4 Summary of Identity and Federation

- Users and Accounts all in AWS
- AWS Organisations
- AWS Control Tower to setup secure and compliant multi-account AWS environment (best practices)
- Federation with SAML
- Federation without SAML with a custom IdP (GetFederationToken)
- IAM Identity Center to connect to multiple AWS Accounts (Organisation) and SAML apps
- Web Identity Federation (not recommended)
- Cognito for most web and mobile applications (has anonymous mode, MFA)
- AWS Directory Service:
  - Managed Microsoft AD - standalone or setup trust AD with on-premises, has MFA, seamless joins, RDS integration
  - AD Connector - proxy requests to on-premises
  - Simple AD - standalone and cheap AD-compatible with no MFA, no advanced capabilities
- AWS RAM to share resource (example VPC subnets)

Chapter 3

Security

3.1 CloudTrail

- Provides governance, compliance and audit for your AWS Account - CloudTrail is enabled by default. - Get a history of events / API calls made within your AWS Account by; - Console - SDK - CLI - AWS Services - Can put logs from CloudTrail into CloudWatch Logs or S3. - A trail can be applied to All Regions (default) or a single Region. - If a resource is deleted in AWS, investigate CloudTrail first

- Management Events: - Operations that are performed on resources in your AWS account - Examples: - Configuring security (IAM AttachRolePolicy) - Configuring rules for routing data (Amazon EC2 CreateSubnet) - Setting up logging (AWS CloudTrail CreateTrail) - By default, trails are configured to log management events. - Can separate Read Events (that don't modify resources) from Write Events (that may modify resources)

- Data Events: - By default, data events are not logged (because high volume operations) - Amazon S3 object-level activity (ex: GetObject, DeleteObject, PutObject): can separate Read and Write Events - AWS Lambda function execution activity (the Invoke API)

- CloudTrail Insights - Enable CloudTrail Insights to detect unusual activity in your account: - Inaccurate resource provisioning - Hitting service limits - Burst of AWS IAM actions - Gaps in periodic maintenance activity - CloudTrail Insights analyses normal management events to create baseline - And then continuously analyses write events to detect unusual patterns - Anomalies appear in the CLoudTrail console - Event is sent to Amazon S3 - An EventBridge event is generated (for automation needs)

- CloudTrail Events Retention - Events are stored for 90 days in CloudTrail - To Keep events beyond this period, log them to S3 and use Athena

3.2 CloudTrail - EventBridge Integration

3.3 CloudTrail - SA Pro

S3 Enhancements: - Enable Versioning - MFA Delete Protection - S3 Lifecycle Policy (S3 IA, Glacier) - S3 Object Lock - SSE-S3 or SSE-KMS encryption - Feature to perform CloudTrail Log File Integrity validation (SHA-256 for hashing and signing)

Observations - The S3 Bucket policy is necessary for cross-account delivery - If Account A wants to access its CloudTrail files: - Option 1: Create a cross-account role and assume the role -

Option 2: edit the bucket policy

- Log filter metrics can be used to detect a high level of API happening - Ex: Count occurrences of EC2 TerminateInstnaces API - Ex: Count of API calls per user - Ex: Detect high level of Denied API calls

- The Organisational Trail is created in the management account.

CloudTrail: How to react to events the fastest? - Overall, CloudTrail may take up to 15 minutes to deliver events - EventBridge: - Can be triggered for any API call in CloudTrail - The fastest, most reactive way. - CloudTrail Delivery in CloudWatch Logs: - Events are streamed - Can perform a metric filter to analyse occurrences and detect anomalies - CloudTrail Delivery in S3 - Events are delivered every 5 minutes - Possibility of analysing logs integrity, deliver cross account, long-term storage.

3.4 KMS

AWS KMS (Key Managment Service) - Anytime you hear ”encryption” for an AWS service, it is most likely KMS - Easy way to control access to your data, AWS manages keys for us. - Fully integrated with IAM for authorisation - Seamlessly integrated into: - Amazon EBS: encrypt volumes - Amazon S3: Server-side encryption of objects - Amazon Redshift: Encryption of data - Amazon RDS: Encryption of data - Etc..... - Can also use the CLI / SDK

KMS - KMS - Key Types - Symmetric (AES-256 keys) - First offering of KMS, single encryption key that is used to Encrypt and Decrypt - AWS services that are integrated with KMS use Symmetric KMS keys - Necessary for envelope encryption - You never get access to the KMS key unencrypted (must call KMS API to use) - Asymmetric (RSA and ECC key pairs) - Public (Encrypt) and Private Key (Decrypt) pair - Used for Encrypt/Decrypt or Sign/Verify operations - The public key is downloadable but you can’t access the Private Key unencrypted - Use case: encryption outside of AWS by users who can’t call the KMS API

Types of KMS Keys - Customer Managed Keys - Create, manage and use. Can enable or disable - Possibility of rotation policy (new key generated every year, old key preserved) - Can add a Key Policy (resource policy) and audit in CloudTrail - Leverage for envelope encryption. - AWS Managed Keys - Used by AWS service (aws/s3, aws/ebs, aws/redshift) - Managed by AWS (automatically rotated every 1 year) - View Key Policy and audit in CloudTrail - AWS Owned Keys - Create and managed by AWS, use by some AWS services to protect your resources - Used in multiple AWS accounts but they are no in your AWS account - You can’t view, use, track or audit.

Types of KMS keys

KMS Key Material Origin - Identifies the source of the key material in the KMS key - Can’t be changed after creation

- KMS(AWS$_KMS)-default-AWSKMScreatedandmanagedthekeymaterialinitsownkeystore$

- External (EXTERNAL) - You import the key material into the KMS key - You’re responsible for securing and managing this key material outside of AWS

- Custom Key Store (AWS$_CLOUDHSM)-AWSKMScreatesthekeymaterialinacustomkeystore(CloudHSM$

KMS Key Source - Custom Key Store (CloudHSM) - Integrate KMS with CloudHSM cluster as a Custom Key Store - Key materials are stored in a CloudHSM cluster that you own and manage - The cryptographic operations are performed in the HSMs - Use cases: - You need direct control over the HSMs - KMS Keys needs to be stored in a dedicated HSMs

KMS Key Store - External - Import your own key material into KMS key, Bring Your Own Key (BYOK) - You’re responsible for key material’s security, availability and durability outside of AWS - Supports both Symmetric and Asymmetric KMS keys - Can’t be used with Custom Key Store

(CloudHSM) - Manually rotate your KMS key (Automatic and On-demand Key Rotation are NOT supported)

KMS Multi-Region Keys

- A set of identical KMS keys in different AWS Regions that can be used interchangeably (same KMS key in multiple Regions) - Encrypt in one Region and decrypt in other Regions (No need to re-encrypt or making cross-Region API calls) - Multi-Region keys have the same key ID, key material, automatic rotation........ - KMS - Multi-Region are NOT global (Primary + Replicas) - Each Multi-Region key is manged independently - Only one primary key at a time, can promote replicas into their own primary - Use cases: Disaster Recovery, Global Data Management (e.g DynamoDB, Global Tables), Active-Active Applications that span multiple Regions, Distributed Signing applicatoins.

3.5 Parameter Store

SSM Parameter Store - Secure storage for configuration and secrets - Optional Seamless Encryption using KMS - Serverless, scalable, durable easy SDK - Version tacking of configurations / secrets - Security through IAM - Notifications with Amazon EventBridge - Integration with CloudFormation

SSM Parameter Store Hierarchy

Standard and advanced parameter tiers

Parameter Policies (for advanced parameters) - Allows to assign a TTL to a parameter (expiration date) to force updating or deleting sensitive data such as passwords. - Can assign multiple policies at a time.

3.6 Secrets Mangager

AWS Secrets Manager - Meant for storing secrets (e.g passwords, API keys,......) - Capability to force rotation of secrets every X days - Automate generation of secrets on rotation (uses Lambda) - Natively supports Amazon RDS (all supported DB engines) - Support other databases and services (custom Lambda function) - Control access to secrets using Resource-based Policy - Integration with other AWS services to natively pull secrets from Secrets Manager: CloudFormation, CodeBuild, ECS, EMR, Fargate, EKS, Parameter Store......

SSM Parameter Store vs Secrets Manager - Secrets Manager (dollah) - Automatic rotation of secrets with AWS lambda - Lambda function is provided for RDS, Redshift, DocumentDB - KMS encryption is mandatory - Can integrate with CloudFormation SSM Parameter Store (dollah) - Simple API - No secret rotation (can enable rotation using Lambda triggered by EventBridge) - KMS encryption is optional - Can integrate with CloudFormation - Can pull a Secrets Manager secret using the SSM parameter Store API

SSM Parameter Store vs Secrets Manager Rotation

3.7 RDS Security

- KMS encryption at rest for underlying EBS volumes / snapshots - Transparent Data Encryption (TDE) for Oracle and SQL Server - SSL Encryption to RDS is possible for all DB (in-flight) - IAM Authentication for MySQL, PostgreSQL and MariaDB - Authorisation still happens within RDS (not in IAM) - Can copy an un-encrypted RDS snapshot into an encrypted one - CloudTrail cannot be used to track queries made within RDS

3.8 SSL Encryption, SNI, MITM

SSL/TLS - Basics - SSL refers to Secure Sockets Layer, used to encrypt connections - TLS refers to Transport Layer Security which is a newer version - Nowadays, TLS certifcates are mainily used but people still refer as SSL - Publis SSL certificates are issued by Certifcate Authorities (CA) - Comodo, Symantec, GoDaddy, GlobalSign, Digicert, Letencrypt, etc - SSL certficates have an expiration date (you set) and must be renewed

SSL Encryption - How it works - Asymmetric Encrutpion is expensive (SSL) - Symmetric encryption is cheaper - Asymmetric handshake is used to exchange a per-client random symmetric key. - Possibilty of client sending an SSL certifccate as well (two-way certificate)

SSL - Server Name Indicattion (SNI) - SNI solves the probelm of loading multiple SSL certificates onto one web server (to server multiple websites) - It's a "newer" protocol and requires the client to indicate the hostname of the target server in the inisital SSL handshake - The server will the find the correct certificate or return the default one.

Note: - Only works for ALB and NLB (newer generation), CloudFront - Does not work for CLB (older gen)

SSL - Man in the middle attacks How to prevent 1 - Don't use public-facing HTTP, use HTTPS (meaning use SSL/TLS certificates) 2 - Use a DNS that has DNSSEC - To send a client to a pirate server, a DNS response needs to be "forged" by a server which intercepts them. - It is possible to protect your domain name by configuring DNSSEC - Amazon Route 53 supports DNSSEC for domain registration - Route 53 supports DNSSEC for DNS service as of December 2020 (using KMS) - You could also run a custom DNS server on Amazon EC2 for example (Bind is the most popular, dnsmasq, KnotDNS, PowerDNS)

3.9 AWS Certificate Manager - ACM

AWS Certificate Manager (ACM) - To host public SSL certificates in AWS, you can - Buy your own and upload them using the CLI - Have ACM provision and renew public SSL certificates for you (free of cost)

- ACM loads SSL certificates on the following integrations: - Load Balancers (including the onces created by EB) - CloudFront distributions - APIs on API gateways

- SSL certificates is overall a pain to manually manage, so ACM is great to leverage in your AWS infrastructure

ACM - Good to know - Possibility of creating public certificates - Must verify public DNS - Must be issued by a trusted public certificate authority (CA) - Possibilty of creating private certificates - For you internal applications - You create your own private CA - Your applications must trust your private CA - Certificate renewal: - Automatically done if generated provisioned by ACM - Any manually uploaded certificates must be renewed manually and re-uploaded. - ACM is a regional service - To use with a global application (multilpe ALB for example), you need to issue an SSL certificate - You cannot copy certs across regions

3.10 CloudHSM

- KMS =¿ AWS manages the software for encryption - CloudHSM -¿ AWS provisions encryption hardware - Dedicated Hardware (HSM = Hardware Security Module) - You manage your own encrypton keys entirely (not AWS) - HSM device is tamper resistant, FIPS 140-2 Level 3 compliance - Supports both symmetric and asymmetric encryption (SSL/TLS keys) - No free tier available - Must use the CloudHSM Client Software - Redshift supports CloudHSM for database encryption and key management - Good option to use with SSE-C encryption

CloudHSM Diagram - IAM permissions - CRUD an HSM Cluster - CloudHSM Software - Manage the Keys - Manage the Users

CloudHSM - High Availability - CloudHSM clusters are spread across Multi AZ (HA) - Great for availability

CloudHSM vs KMS

3.11 Solution Architecture - SSL on ELB

- You can offload SSL to CloudHSM (SSL Acceleration) - Supported by NGINX Apache Web servers and IIS for windows server - Extra security: the SSL private key never leaves the HSM device - Must setup a cryptographic user (CU) on the CloudHSM device

3.12 S3 Security

- SSE-S3: encrypts S3 objects using keys handled and managed by AWS - SSE-KMS: leverage KMS to manage encryption keys - Key usage appears in CloudTrail - objects made public can never be read - On s3:PutObject make the permission kms:GenerateDataKey is allowed - SSE-C: when you want to manage your own encryption keys - Client-Side Encryption

Glacier: all data is AES-256 encrypted, key under AWS control

Encryption in transit (SSL/TLS) - Amazon S3 exposes: - HTTP endpoint: non encrypted - HTTPS endpoint: encryption

- You're free to use the endpoint you want, but HTTPS is reccomended - HTTPS is mandatory for SSE-C - To enforce HTTPS, use a bucket policy wit aws:SecureTransport

Events in S3 Buckets - S3 Access Logs: - Detailed records for the reqeusts that are made to a bucket - Might take hours to deliver - Might be incomplete (best effoirt) - S3 Event Notifications - Recieve notifications when certain events happen in your bucket - E.g: new objects created, object removal, restore objects, replication events. - Destinsations: SNS, SQS queue, Lambda - Typically delivered in seconds but can take minutes, notification for every object if versioning is enabled, else risk of one notification for two same objects write done simultaneously. - Trusted Advisor: - Check the bucket permission (is the bucket public?) - Amazon EventBridge - Need to enable CloudTrail object level logging on S3 first - Target can be Lambda, SQS, SNS, etc.........

S3 Security

- User based - IAM policies - which API calls shoudl be allowed for a specific user from IAM console

- Resource Based - Bucket Policies - bucket wide rules from the S3 console - allows cross account - Object Access Contorl List (ACL) - finer grain - Bucket Access Contorl List (ACL) - less common

S3 Bucket Policies - Use S3 bucket for policy to: - Grant public access to the bucket - Force objects to be encrypted at uplade - Grant access to another account (Cross Account) - Optional Conditions on: - SourceIp: Public IP or Elastic IP — VpcSourceIp: Private IP (through VPC Endpoint) - Source VPC or Source VPC Endpoint - only workds with VPC Endpoints - CloudFront Origin Identity - MFA

S3 pre-signed URLs - Can generate pre-signed URLs using SDK or CLI - For downloads (easy can use the CLI) - For uploads (harder must use the SDK) - Valid for a default of 3600 seconds, can change timeout with –expires-in[TIME$_BY_SECONDS]argument-Usersgivenapre-signedURLinheritthepermissionsofthepersonwhogeneratedtheURLforGET/PUT$

Examples: - Allow only logged-in users to download a premium video on your S3 bucket - Allow an ever changing list of users to download files by generating URLs dynamically - Allow temporarily a user to upload a file to a precise location in our bucket

VPC Endpoint Gateway for S3

S3 Object Lock and Glacier Vault Lock - S3 Object Lock - Adopt a WORM (Write once read many) model - Block an object version deletion for a specified amount of time - Glacier Vault Lock - Adopt a WORM (Write Once Read Many) model - Locak the policy for future edits (can no longer be changed) - Helpful for compliance and data retention

3.13 S3 Access Points

- Access Points simplify security managment for S3 buckets - Each access point has: - its own DNS name (Internet Origin or VPC origin) - an access point policy (similar to bucket policy) - manage security at scale

S3 - Access Points - VPC Origin - We can define the access point ot be accessible only form within the VPC - You must create a VPC Endpoint to access the Access Point (Gateway or Interface Endpoint) - The VPC Endpoint Policy must allow access to the target bucket and Access Point

3.14 S3 Multi-Region Access Points

- Provide a global endpoint that spans S3 buckets in multiple AWS regions - Dyanamically route requests to the nearest S3 bucket (lowest latency) - Bi-directional S3 bucket replication rules are created to keep data in sync across regions - Failover Contorls - allows you to shift request across S3 buckets in different AWS regions within minutes (Active-Active or Active-Passive)

Multi-Region Access Points - Failover Controls

3.15 S3 Multi-Region Access Points - Hands On

3.16 S3 Object Lambda

- Use AWS Lambda Functions to change the object before it is retrieved by the caller application - Only one S3 bucket is needed on top of whcih we create S3 Access Point and S3 Object Lambda Access Points - Use Cases: - Redacting personally indentifiable information for analytics or non production environments - Converting across data formats, such as converting XML to JSON - Resizing and watermarking images on the fly using caller-specific details, such as the user who requrest the object

3.17 DDoS and AWS Shield

What is a DDoS

Types of Attacks on your infrastrcuture - Distributed Denial os Servicve (DDoS) - When your service is unavailable because it is receiving too many requests - SYN Flood (Layer 4) - send too many TCP connection reqeusts - UDP reflection (Layer 4) - get other servers to send many big UDP requests - DNS flood attack: Overwhelm the DNS so legitimate users can't find the site - Slow Loris attack: a lot of HTTP connections are opened and maintained (hmmmm consider websocket servers!!)

- Application Level Attacks: - more complext, more specific (HTTP level) - Cache bursting strategies: overload the backend database by invalidating the cache (clever)

DDoS Protection on AWS - AWS Shield Standard: Protects against DDoS attack for your website and applications, for all customers at no additional cost - AWS Shield Advanced: 24/7 premium DDoS protection - AWS WAF: Filter specific requests based on rules - CloudFront and Route 53: - Availability protection using global edge network - Combined with AWS shield: provides DDoS attack mitigation at the edge - Be ready to scale - leverage AWS Auto Scaling - Separate static resources (S3 / CloudFront) form dynamic ones (EC2 / ALB) - Read the whitepaper

Sample Reference Architecture

AWS Shield - AWS Shield Standard - Free service that is activated for every AWS customer - Provides protection from attacks such SYN/UDP Floods, Reflection attacks and other layer 3/layer 4 attacks - AWS Shield Advanced - Optional DDoS mitigation service (3000 dollahs per month per organisation) - Protect against more sophisticated attack on Amazon EC2, Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, Route 53 - 24/7 access to AWS DDoS response team (DRP) - Protect against higher fees during usage spikes due to DDoS

3.18 AWS WAF - Web Application Firewall

- Protects your web applications from common web exploits (Layer 7) - Deploy on Application Load Balancer (localised rules) - Deploy on API Gateway (ruels running at the regional or edge level) - Deploy on CloudFront (rules on edge locations) - Used in front of other solutions: CLB, EC2 Instances, custom origins, S3 websites - Deploy on AppSync (protect your GraphQL APIs) - WAF is not for DDoS protection - Define Web ACL (Web Acess Control List) - Rules can include IP addresses, HTTP headers, HTTP body, or URI strings - Protects from common attack - SQL injection and Cross-Site scripting (XSS) - Size constaints, Geo Match - Rate-based rules (to count occurrences of events) - Rule Actions: Count — Allow — Block — CAPTCHA — Challenge

AWS WAF - Mangaged Rules - Library of over 190 managed ruels - Ready to use ruels that are managed by AWS and AWS Marketplace Sellers

- Baseline Rule Groups - general protection from common threats - AWSManagedRulesCommonRuleSet, AWSManagedRulesAdminProtectionRuleSet,..... - Use-case Specific Rule Groups - Protection for many AWS WAF use cases - AWSManagedRulesSQLLiRuleSet , AWSManagedRulesWindowsRulesSet, AWSMangedRulesPHPRuleSet, AWSManagedRulesWordPressRuleSet - IP Reputation Rule Groups - block requests based on source (e.g. malicious IPs) - AWSManagedRulesAmazonIpReputationList, AWSManagedRulesAnonomoyusIpList - Bot Control Managed Rule Group - block and manage requests from bots - AWSMangedRulesBotContorlRuleSet

WAF - Web ACL - Logging - You can send your logs to an: - Amazon CloudWatch Logs log group - 5 MB per second - Amazon Simple Storage Service (Amazon S3) bucket - 5 minutes interval - Amazon Kinesis Data Firehouse - limited by Firehose quotas

Solution Architecture - Enhance CloudFront Origin Security with AWS WAF and AWS Secrets Manager

3.19 AWS Firewall Manager

- Manage rules in all accounts of an AWS organisation - Security policy: common set of security rules - WAF rules (Application Load Balancer, API Gateways, CloudFront) - AWS Shield Advanced (ALB, CLB, NLB, Elastic IP, CloudFront) - Security Groups for EC2, Application Load Balancer and ENI resources in VPC - AWS Network Firewall (VPC Level) - Amazon Route 53 Resolver DNS Firewall - Policies are created at the region level

- Rules are applied to new resources as they are created (good for compliance) across all and future accounts in your Organisation

WAF vs Firewall Manager vs Shield - WAF, Shield and Firewall Manager are used together for comprehensive protection - Define your Web ACL rules in WAF - For granular protection of your resources, WAF along is the correct choice - If you want to use AWS WAF across accounts, accelerate WAF configuration, automate the protection of new resources use Firewall Manager with AWS WAF - Shield Advanced adds additional features on top of AWS WAF, such as dedicated support from the shield response team (SRT) and advanved reporting - If you're prone to frequent DDoS attacks, consider purchasing Shield Advanved

3.20 Blocking an IP Address

Blocking an IP address

Blocking an IP address - with an ALB

Blocking an IP address - with an NLB

Blocking an IP address - ALB + WAF

Blocking an IP address - ALB, CloudFront and WAF

3.21 Amazon Inspector

Amazon Inspector - Automated Security Assessments - For EC2 instances - Leveraging the AWS System Manager (SSM) agent - Analyse against unintended network accessibility - Analyse the running OS against known vulnerabilities - For container images push to amazon ECR - Assessment of Container Images as they are pushed - For Lambda Functions - Identifies software vulnerabilities in function code and package dependencies - Assessment of functions as they are deployed

- Reporting and integration with AWS security hub - Send findings to Amazon Event Bridge

What does Amazon Inspector evaluate? - Remember: only for EC2 instances, Container Images and Lamabda functions - Continuous scanning of the infrastructure, only when needed - Package vulnerabilities (EC2, ECR and Lambda) - database of CVE - Network reachability (EC2) - A risk score is associated with all vulnerabilities for prioritisation

3.22 AWS Config

AWS COnfig - Help with auduting and recording compliance of your AWS resources - Helps records configuration and changes over time - AWS Conifig rules does not prevent actions from happening (no deny) - Questions that can be solved by AWS config - Is there unrestricted SSH access to my security groups? - Do my buckets have any public access? - How has my ALB configuration changed over time? - You can recieve alerts (SNS notifications) for any changes - AWS Config is a per-region service - Can be aggregated across regions and accounts

AWS config resource - View compliance of a resource over time

- View configuration of a resource over time

- View CloudTrail API calls if enabled

AWS Config Rules - Can use AWS managed config rules (over 75) - Can make custom config rules (must be defined in AWS Lambda) - Evaluate if each EBS disk is of type gp2 - Evaulate if each EC2 instance is t2.micro - Rules can be evaluated / triggered - For each config change - And / or: at regular time intervals - Trigger Amazon EventBridge if the rule is non-compliant (chain with Lambda) - Rules can have auto remediations through SSM Automations - If a resource is not compliant you can trigger an auto remediation - Ex: remediate security group rules, stop instances with non-approved tags

3.23 AWS Managed Logs

- Load Balancer Access Logs (ALB, NLB, CLB) -¿ S3 - Access logs for your Load Balancers - CloudTrail Logs -¿ to S3 and CloudWatch logs - Logs for API calls made within your account - VPC Flow Logs -¿ to S3, CloudWatch Logs, Kinesis Data Firehose - Information about IP traffic going to and from network interfaces in your VPC - Route 53 Access Logs -¿ to CloudWatch logs - Log information about the queries that Route 53 receives - S3 Access Logs -¿ to S3 - Server access logging provides detailed records for the requests that are made to a bucket - CloudFront Access Logs -¿ to S3 - Detailed information about every user request that CloudFront receives - AWS Config -¿ to S3

3.24 Amazon GuardDuty

- Intelligent Threat discovery to protect your AWS Account - Using Machine Learning algorithms, anomaly detection, 3rd party data - One click to enable (30 days trial) no need to install software - Input data includes: - CloudTrail Events Logs - unusual API calls, unauthrosed deployments - CloudTrail Management Events - create VPC subnet, create trail - CloudTrail S3 Data Events - get object, list objects, delete object - VPC Flow logs - unusual internal traffic, unusual IP address - DNS Logs - compromised EC2 instances sending encoded data within DNS queries - Optional Feature - EKS Audit Logs, RDS and Aurora, EBS, Lambad, S3 Data Events....... - Can setup EventBridge rules to be notified in case of findings - EventBridge rules can target AWS Lambda or SNS - Can protect against CryptoCurrency attacks (has a dedicated ”finding” for it)

Amazon GuardDuty

GuardDuty - Delegated Administrator - AWS organisation member accounts can be designated to be a GuardDuty Delegated Administrator - Have full permissions to enable and manage GuardDuty for all accounts in the Organisation - Can be done only using the Organisation Mangement Account

3.25 IAM Advanced Policies

IAM conditions

IAM for S3 - s3:ListBucket permission applies to arn:aws:s3:::test -¿ bucket level permission

- s3:GetObject, s3:PutObject,s3:DeleteObject applies to arn:awn:s3:::test - Object level permissions

Resource Policies and aws:PrincipalOrgID - aws:PrincipalOrgID can be used in any resource polices to restrict access to accounts that are memeber of an AWS organisation

3.26 EC2 Instance Connect

EC2 Instance Connect (SendSSHPublicKey API)

3.27 AWS Security Hub

- Central security tool to manage security across several AWS accounts and automate security checks - Integrated dashboards showing current security and compliance status to quickly take actions - Automatically aggregates alerts in predefined or personal findings formates from various AWS services and AWS partner tools: - Config - GuardDuty - Inspector - Macie - IAM Access Analyser - AWS Systems Manager - AWS Firewal Manager - AWS Health - AWS Partner Network Solutions

Must first enable the AWS config service

3.28 Amazon Detective

- GuardDuty, Macie and Security Hub are used to identify potential security issues or findings. - Sometimes security findings require deeper analysis to isolate the root cause and take actions - it is a complext process - Amazon Detective analyses, investigates and quickly identifies the root cause of security issues or suspicious activities (using ML and graphs) - Automatically collects and processes events from VPC Flow Logs, CloudTrail, GuardDuty and creates a unifed view. - Produces visualisations with details and context to get to the root cause.

Chapter 4

Compute & Load Balancing

4.1 Solution Architecture on AWS

- DNS Layer - Route 53 - Web Layer - CLB, ALB, NLB, API Gateway, Elastic IP - Compute Layer - EC2, ASG, Lambda, ECS, Fargate, Batch, EMR - CDN Layer - CloudFront - Caching / Session Layer - ElastiCache, DAX, DynamoDB, RDS - Database Layer - RDS, Aurora, DynamoDB, ElastiSearch, S3, Redshift - Decoupling Orchestration Layer - SQS, SNS, Kinesis, Amazon MQ, Step Functions - Storage Layer - EBS, EFS, Instance Store - Static Assets Layer (storage) - S3, Glacier

4.2 EC2

EC2 Instance Types - Main Ones - R: applications that need a long of RAM - in-memory caches - C: applicatoins that need good CPU - compute / databases - M: Applications that are balanced (think ”medium”) - general / web app - I: Aplications that need good local I/O (instance storage) - databases - G: Applications that need a GPU - video rendering / machine learniing - T2 / T3: burstable instances (up to a capacity) - T2 / T3 - unlimited: unlimited burst

See link at https://www.ec2instances.info.

EC2 - Placement Groups - Control the EC2 instance placement strategy using placement groups - Group strategies - Cluster - cluster instances into a low-latency group in a single Avialability Zone - Spread - spreads instances across undelying hardward (max 7 instances per group per AZ) - critical applications - Partition - spreads instances across many different partitions (which rely on different sets of racks) within an AZ. Scales to 100s of EC2 instances per group (Hadoop, Cassandra, Kafka) - You can move an instance into or out of a placement group - You first need to stop it - You then need to use the CLI (modify-instance placement) - You can then start your instances

Placement Groups Cluster - Pros: Greate network (10 Gbps bandwidth between instances with Enhanced Networking enabled - recommended) - Cons: If the rack fails, all instances fails at the same time. - Use case: - Big Data job that needs to complete fast - Applicatoins that needs extremely low latency and high network throughput

Placement Groups Spread - Pros - Can span across Availability Zones (AZ) - Reduced risk of simulataneous failure - EC2 Instances are on different pysical hardware - Cons - Limited to 7 instances per AZ per placement group - Use case: - Application that need to maximise high availability - Critical Applications where each instance must be isolated from failure from each other

Placement Groups Partition - Up to 7 partitions per AZ - Up to 100s of EC2 instances - The instances in a partition do not share racks with the instances in the other partitions - A partition failure can affect many EC2 instances but won't affect other partitions - EC2 instances get access to the partition information as metadata - Use cases: HDFS, HBase, Cassandra, Kafka

EC2 Instance Launch Types - On Demand Instances: Short workload, prefictable pricing, reliable - Spot Instance: short workloads, for cheap, can lose instances (not reliable) - Reserved (Minimum 1 year) - Reserved Instances: Long workloads - Convertible Reserved Instances: Long workloads with flexible instances - Highest to lowest discount: All upfront payment, partial upfront payment, no upfront - Dedicated Instances: No other customers will share your hardware - Dedicated Hosts: book an entire physical server, control instance placement - Great for software licenses that operate at the core, or CPU socket level - Can define host affinity so that instance reboots are kept on the same host

EC2 Graviton - AWS Graviton Processors deliver the best price performance - Supports many Linux OS, Amazon Linuz 2, RedHat, SUSE, Ubuntu - Not available for windows instances - Graviton2 - 40% better price performance over comparable 5th generation x86 based instances - Graviton3 - Up to 3x better perfromance compared to Graviton2 - Use cases: app servers, microservices, HPC, CPU-based ML, video encodign, gaming, in-memory cahces,.....

EC2 included metrics - CPU: CPU Utilisations + Credit Usage / Balance - Network: Network In / Out - Status Check - Instance status = check the ECM VM - System status = check the undelying hardware - Disk: Read / Write for Ops / Bytes (only for instance store) - RAM is not included in the AWS EC2 metrics

EC2 Instance Recovery - Status Check - Instance Status = check the EC2 VM - System status = check the underlying hardware - Recovery: Same Private, Public, Elastic IP, metadata, placement group

4.3 High Performance Computing (HPC)

- Cloud is good for HPC - Can create a very high number of resources in no time. - Can speed up time to results by adding more resources - You can pay only for the systems you have used —- Examples: Genomics, computational chemistry, financial risk modelling, weather prediction, machine learning, deep learning, autonomous driving...... - Services which help HPC?

Data Management % Transfer - Aws Direct Connect - Move GBs of data to the cloud over a private secure network - Snowball - Move PB of data to the cloud - AWS DataSync - Move large amount of data between on-premise and S3, EFS, FSx for Windows

Computer and Networking - EC2 Instances: - CPU optimised, GPU optimised - Spot Instances / Spot Fleets for cost savings + Auto Scaling - EC2 Place Groups: Cluster for good network performance - EC2 Enhanced Networkin (SR-IOV) - Higher Bandwidth, higher PPS (packet per second), lower latency - Option 1: Elastic Network Adapted (ENA) up to 100 Gbps - Option 2: Intel 82599 up to 10Gbps - Legacy - Elastic Fabric Adapter (EFA) - Improved ENA for HPC only works for Linux - Great for inter-node communications, tightly coupled workloads - Leverages Message Passing interface (MPI) standard - Bypasses the undelying Linux OS to provide low-latency reliable transport

Storage - Instance-attached storage - EBS: scale up to 256000 IOPS with io2 Block Express - Instance Store: Scal to milliiton of IOPS, linkedin to EC2 instance, low latency.

- Network Storage - Amazon S3: large blob, not a file system - Amazon EFS: scale IOPS based on total size or use provisioned IOPS - Amazon FSx for Lustre: - HPC optimised distributed file system, millions of IOPS - Backed by S3

Automation and Orchestration - AWS Batch - AWS Batch supports multi-node paralell jobs which enables you to run single jobs that span multiple EC2 instances - Easily schedule jobs and launch EC2 instances accordingly - AWS ParalellCluster - Open source cluster management tool to deploy HPC on AWS - Configure text files - Automate creation of VPC, Subnet, cluster type and instance types

4.4 Auto Scaling

Auto Scaling Groups - Dynamic Scaling Policies - Target Tracking Scaling - Most simple and easy to setup - Example: I want the average ASG CPU to stay at aroudnd 40% - Simple / Step Scaling - When a CloudWatch alarm is triggered (example CPU ¿ 70%) then add 2 units - When a CloudWatch alarm is triggered (example CPU ¡ 30%) then remove 1 - Sceduled Actions - Anticipate a scaling based on known usage patterns - Example: Increase the min capcity to 10 at 5pm on Fridays

Auto Scaling Groups - Predictive Scaling - Predictive Scaling: continuously forcast load and scedule scaling ahead

Good metrics to scale on - CPUUtilisation: Average CPU utilisation across your instances - RequstCountPerTarget: to make sure the number of reqeusts per EC2 instances is stable - Average Network In / Out (if you're application is network bound) - Any custom metric (that you push using CloudWatch)

Auto Scaling - Good to know - Spot fleet support (mix on Spot and On-Demand instances) - Lifecycle Hooks - Perform actions before an instance is in service, or before it is terminated - Example: Cleanup, log extraction, special health checks - To upgrade an AMI, must update the launch configuration / template - Then terminate instances manually (CloudFromation can help) - Or use EC2 instance refresh for Auto Scaling

Auto Scaling - Instance Refresh - Goal: Update launch template and then re-creating all EC2 instances - For this we can use the native feature of Instance Refresh - Setting of minimum healthy percentage - Specify warm-up time (how long until the instance is ready to use)

Auto Scaling - Scaling Processes - Launch: Add a new EC2 to the group, increasing the capacity - Terminate: Remove an EC2 instance from the group, decreasing the capcity - HealthCheck: Check the health of the instances - ReplaceUnhealty: Terminate unhealthy instances and re-create them - AZRebalance: Balancer the number of EC2 instances across AZ - Alarm Notification: Accept notification from CloudWatch - SchedulesActions: Performs scheduled actions that you create - AddToLoadBalancer: Add instances to the load balancer or target group. - InstanceRefresh: Perform an instance refresh —-¿ These processes can be suspended

Auto Scaling - Health Checks - Health Checks available: - EC2 Status Checks - ELB Health Checks (HTTP) - Customer Health Checks - send instances health to an ASG using AWS CLI or AWS SDK (set-instnance-health) - ASG will launch a new instance after terminating an unhealthy one - Make sure the health check is simple and check ths correct thing

4.5 Auto Scaling Update Strategies

Auto Scaling - Updating an application

Auto Scaling - Solution Architecture

4.6 Spot Instances and Spot Fleet

EC2 Spot Instances - Can get a discount of up to 90% compared to On-Demand - Define max spot price and get the instance while current spot price ¡ max - The hourly spot price variaes based on offer and capacity - If the current spot price ¿ your max price you can choose to stop or terminate your instance with a two minutes grace period - Used for batch jobs, data analysis or workloads that are resilient to failures. —¿ Not great critical jobs or databases

Spot Fleets - Spot Fleets = set of Spot Instances + (optional) On-Demand Instances - The Spot Fleet will try to meet the target capacity with price constraints - Define possible launch pools: instance type (m5.large), OS, Availability Zone - Can have mulltiple launch pools, so that the flee can choose - Spot fleets stops launching instances when reaching capacity or max cost. - Strategies to allocate spot instances - lowestPrice: form the pool with the lowest price (cost optimisation, short workload) - diversified: distribute across all pools (great for availability, long workloads) - capcityOptimised: pool with optimal capacity for the number of instances - priceCapacityOptimised (recommended): pools with highest capacity available then select the pool with the lowest price (best choice for most workloads) –¿ Spot fleets allow us to automaticaly requests spot instances with the lowest price

4.7 Amazon ECS - Elastic Container Service

What is Docker? - Docker is a software development platfform to deploy appls - Apps are packaged in containers that can be run on any OS - Apps run the saem, regardless of where they are run - Any machine (no compatibilty issues, predictable behaviour) - Less work - Easier to maintain and deploy - Works with any language, any OS, any technology. - Control how much memeory / CPU is aollcated to your continers - Scale containers up and down very quickly (seconds) - More effecient than virtual machines

Docker Containers Management on AWS - To manage containers, we need a container management platform - Amazon Elastic Container Service (Amazon ECS) - Amazon's own container platform - Amazon Elastic Kubernetes Service (Amazon EKS) - Amazon managed Kubernetes (open source) - AWS Fargate - Amazons own Serverless container platform - Works with ECS and with EKS

Amazon ECS - Use cases - Run Microservices - Run multiple Docker containers on the same machine - Easy service discovery features to enhance communication - Direct integration with Application Load Balancer and Network Load Balancer - Auto Scaling capability - Run Batch Processing / Scheduled Tasks - Scehdule ECS tasks to run on On-Demand / Reserved / Spot Instances - Migrate Applications to the Cloud - Dockerise legacy applications runngin on-premisses - Move docker containers to run on Amazon ECS

Amazon ECS - Concepts - ECS Cluster - logical grouping of EC2 instances - ECS Service - defines how many tasks shoudl run and how they should be run - Task definitions - metadata in JSON form to tell ECS how to run a Docker containers (image naem, CPU, RAM) - ECS Task - an instance of a Task Definition, a running Docker containers - ECS IAM roles - EC2 Instance Profile - used by the EC2 instance (e.g make API calls to ECS, send logs) - ECSTask IAM Role - allow each task to have a specific role (e.g make API calls to S3, DynamoDB)

Amazon ECS - ALB Integration - We get Dynamic Port Mapping - Allows you to run multiple instances of the same application on the same EC2 instance - The ALB finds the right port on your EC2 instances - Use cases: - Increases resiliency even if running on one EC2 instance - Maximise ustilisation of CPU cores - Ability to perform rolling upgrades withouth impacting app uptime

AWS Fargate - Launch Docker containers on AWS - You do not provision the infrastructure (no EC2 instances to manage) - Its all serverless - You create task definitions - AWS run containers for you based on the CPU / RAM you need - To scale, just increase the number of tasks——no more EC2 instances

Amazon ECS - Security and Netoworking - You can inject secrets and configurations as Environment Variable into running Docker containers - Integration with SSM parameter store and secrets manager - ECS Tasks Networking - none - no network connectivity, no port mappings - bridge - uses Docker's virtual container based network - host - bypass Docker's network uses the underlying host network interface - awsvpc - Every task launched on the instance gets it own ENI and a private IP adress - Simplified networking, enhanced security, Security Groups, monitoring, VPC Flow Logs - Default mode for Fargate tasks

Amazon ECS - Service Auto Scaling - Automaticalyly increase / decrease the desired number of tasks - Amazon ECS leverages AWS application auto scaling - CPU and RAM is tracked in CloudWatch as the ECS Service level - Target Tracking - scale based on target value for a specific CloudWatch metrics - Step Scaling - scale based on a specified CloudWatch Alarm - Scheduled Scaling - scale based on a specified date/time (predictable changes) - ECS Service Auto Scaling (task level) $\neq$ EC2 Auto Scaling (EC2 instance level) - Fargate Auto Scaling is much easier to setup (because Serverless)

Amazon ECS - Spot Instances - ECS Classic (EC2 Launch Type) - Can have the underlying EC2 instances as Spot Instances (managed by an ASG) - Instances may go into draining mode to remove running tasks - Good for cost savings, but will impact reliabilty - AWS Fargate - Specify minimum of tasks for on-demand baseline workload - Add tasks running on FARGATE_SPOT for cost-savings (can be reclaimed by AWS) - Regardless of On-demand or Spot, Fargate scales well based on load

4.8 Amazon ECR - Elastic Container Registry

- Store and manage Docker images on AWS - Private and Public repository (Amazon ECR Public Gallery https://gallery.ecr.aws) - Fully integration with ECS - Access is controlled through IAM (permission errors =¿ check policy) - Supports image vulnerabilty scanning versioning, image tags, image lifecycle.....

Amazon ECR - Cross Region Replications - ECR Private registry supports both cross-Region and cross-account replication

Amazon ECR - Image Scanning - Manual Scan or Scan on Push - Basic Scanning - Common CVE - Enhanced Scanning - Leverages Amazon Inspector (OS and Programming Language vulnerabilties) - Scan results can be retrieved from within the AWS console

4.9 Amazon EKS - Elastic Kubernetes Service

- Amazon EKS = Amazon Elastic Kubernete Service - It is a way to launch managed Kubernets clusters on AWS - Kubernetes is an open-source system for automatic deployment, scaling and management of containerised (usually docker) application - It is an alternative to ECS, similar goal but different API - EKS supprots EC2 if you want to deploy worker nodes or Fargate to deploy serverless containers - Use Case: If you company is already using Kubernetes on-premises or in another cloud and wants to migration to AWS using Kubernetes —-¿ Kubernetes is cloud-agnostic (can be used in any cloud - Azure, GCP) - For multiple regions, deploy on EKS cluster per region. - Collects logs and metrics using CloudWatch Container Insights

Amazon EKS - Diagram

Amazon EKS - Node Types - Managed Node Groups - Creates and manages Nodes (EC2 Instances) for you - Nodes are part of an ASG managed by EKS - Supports On-Demand or Spot Instances - Self-Managed Nodes - Nodes created by you an registered to the EKS cluster and managed by an ASH - You can use prebuilt AMI - Amazon EKS optimised AMI - Supports On-Demand or Spot Instances - AWS Fargate - No maintenance required; no nodes managed.

Amazon EKS - Data Volumes - Need to specify StorageClass manifest on your EKS cluster - Leverages a Container Storage Interface (CSI) compliant driver - Support - Amazon EBS - Amazon EFS (works with Fargate) - Amazon FSx for Lustre - Amazon FSx for NetApp ONTAP

4.10 AWS App Runner

- Fully managed service that makes it easy to deploy web applications and APIs at scale - No infratructure experience requried - Start with your source code or container image - Automatically builds and deploy the web app - Automatic scaling highly available, load balancer, encryption - VPC access support - Connect to database, cache and message queue services - Use cases: web apps, APIs, microservices, rapid production deployments

Solution Architecture - App Runner Multi-Region Architecture

4.11 ECS Anywhere and EKS Anywhere

Amazon ECS Anywhere - Easily run containers on Customer-managed infrastructure (on-premises, VMs) - Allows customers to deploy native Amazon ECS tasks in any environment - Fully-manged Amazon ECS Control Plane - ECS Container Agent and SSM Agent needs to be installed - ËXTERNALL̈aunch Type —-¿ Must have a stable connectoin to the AWS region - Use cases: - Meet compliance, regulatory and latency requirements - Run apps outside AWS regions and closer to their other services - On-premises ML, video processing, data processing.

Amazon EKS Anywhere - Create an operate kubernetes clusters created outside AWS - Leverage the Amazon EKS Distro (AWS' bundled release of Kubernetes) - Reduce support costs and avoid maintaining redundant thrid party tools. - Install using the EKS Anywhere installer - Optionally use the EKS Connector to connect the EKS Anywhere clusters to AWS - Full Connected and Partially Disconnected: you can connect to Amazon EKS Anywhere clusters to AWS and leverage the EKS console - Fully Disconnected: Must install the EKS Distro and leverage open-source tools to manage your clusters.

4.12 AWS Lambda - Part 1

AWS Lambda Integration Main ones - API Gateway - Kinesis - DynamoDB - AWS S3 - Simple Storage Service - AWS IoT - Internet of Things - Amazon EventBridge - CloudWatch Logs - AWS SNS - AWS Cognito - Amazon SQS

AWS Lambda Language Support (runtimes) - node.js (javascript) - Python - Java - C (.NET Core) / Powershell - Ruby - Custom Runtime API (community supported, example Rust or Golang) hmmmmmmmm.......... - Lambda Container Image - The container image must implement the Lambda Runtime API - ECS / Fargate is preferred for running arbitrary Docker images

Lambda - Limits to know – RAM - 128 to 10,240 MB (10 GB) – CPU - is linked to RAM (cannot be set manually) - 2 vCPUs are allocated at 1,769 MB of RAM - 6 vCPUs are alloacted at 10,240 MB of RAM - Timeout - up to 15 minutes - /tmp Storage - 10,240 MB - Deployment Package - 50 MB (zipped), 250 MB (unzipped) including layers - Concurrent Executions - 1000 (soft limit that can be increased) - Container Image Size - 10GB - Invocation Payload (reqeust/ response) - 6 MB (sync), 256KB (async)

Lambda Concurrency and Throttling - Concurrency limit: up to 1000 concurrent executions - Can set a 'reserved concurrency' at the funciton level (=limit) - Each invocation over the concurrency limit will trigger a "Throttle" - Can request a quote increase in AWS Service Quotes

Lambda Concurrency Issue - If you don't reserve (=limit) concurrecny, the following can happen

Lambda and CodeDeploy - CodeDeploy can help you automate traffic shift for Lambda aliases - Feature is integrated within the SAM framework - Linear: grow traffic every N minutes until 100% - Linear10PercentEvery3Minutes - Linear10PercentEvery10Minutes - Canary: try X percent then 100% - Canary10Percent5Mintues - Canary10Percent30Minutes - AllAtOnce: immediate - Can create Pre and Post Traffic Hooks to check the health of the Lambda function

AWS Lambda Logging, Monitoring and Tracing - CloudWatch: - AWS Lamabda execution logs are stored in AWS CloudWatch Logs - AWS Lambda metrics are displayed in AWS CloudWatch Metrics (successful invocations, error rates, latency, timeouts, etc....) - Make sure you AWS Lambda functions has an execution role with an IAM policy that authorises writes to CloudWatch Logs - X-Ray - It's possible to trace Lambda with X-Ray - Enable in Lambda configurations (runs the X-Ray daemon for you) - Use AWS SDK in Code - Ensure Lambda Function has correct IAM Execution Role

4.13 AWS Lambda - Part 2

Lambda in a VPC Note: Lambda: CloudWatch Logs works even withouth endpoint or NAT gateway

Lambda - Fixed Public IP for external comms

Lambda - Synchronous Invocations - Synchronous: CLI, SDK, API Gateway - Results is returned right away - Error handling must happen client side (retries, exponential backoff, etc)

Lambda - Asynchornous Invocation - S3, SNS, Amazon EventBridge - Lambda attempts to retry on errors (3 tries total) - Make sure the processing is idempotent (in case of retries) - Can define a DLQ (dead letter queue) - SNS or SQS for failed processing

Lambda - Architecture Discussion - Start immediately paralell executions - Amazon S3 -¿ Amazon SNS -¿ Lambda - Batched Executions Delay - Amazon S3 -¿ Amazon SNS -¿ Amazon SQS -¿ Lambda

4.14 Elastic Load Balancers - Part 1

Types of load balancer on AWS - AWS has 4 kinds of managed Load Balancers - Classic Load Balancer (v1 - old generation) - 2009 - CLB - HTTP, HTTPS, TCP, SSL (secure TCP) - Application Load Balancer (v2 - new generation) - 2016 - ALB HTTP, HTTPS, WebSocket - Network Load Balancer (v2 - new generation - 2017 - NLB) - TCP, TLS (secure TCP), UDP - Gateway Load Balancer - 2020 - GWLB - Operates a layer 3 (Network Layer) - IP Protocol

- Overall, it is reccomended to use the newer generation load balancers as they provide more features - Some load balancers can be setup as internal (private) or external (public) ELBs

Classic Load Balancers (v1) - Heath Checks can be HTTP (L7) or TCP (L4) based including with SSL - Supports only one SSL certificate - The SSL certificate can have many SAN (Subject Alternate Name), but the SSL cerficate must be changed anytime a SAN is added / edited / removed - Better to use ALB with SNI (Server Name Indication) is possible - Can use multiple CLB if you want distinct SSL certifcates - TCP -¿ TCP passes all the traffic to the EC2 instance - Only way to use 2-way SSL authetication

Application Load Balancer (v2) - Application Load Balancers is Layer 7 (HTTP) - Load balancing to multiple HTTP applications across machines (target groups) - Load balancing to multiple applications on the same machine (ex: containers) - great fit with ECS and has dynamic port mapping - Support for HTTP/2 and WebSocket - Support redirects (from HTTP to HTTPS for example) - Routing Rules for path, headers, query string.

Application Load Balancer (v2) - HTTP Based Traffic

Application Load Balancer (v2) - Target Groups - EC2 Instances (can be managed by an Auto Scaling Group) - HTTP - ECS tasks (managed by ECS itself) - HTTP - Lambda functions - HTTP request is translated into a JSON event - IP Addresses - must be private IPs - ALB can route to multiple target groups - Health Checks are at the target group level

Network Load Balancer (v2) - Network load balancers (Layer 4) allow to: - Forward TCP and UDP traffic to your instances - Handles milliions of request per seconds - Less latency (100 ms) (vs 400ms for ALB)

- NLB has one static IP per AZ and supports assigning Elastic IP (helpful for whitelising specific IP) - NLB are used for extreme performance, TCP or UDP traffic - Not included in the AWS free tier

Network Load Balancer - Target Groups - EC2 instances - IP Addresses - must be private IPs - Application Load Balancer

Network Load Balancer – Zonal DNS Name - Resolving Regional NLB DNS name returns the IP addresses for all NLB nodes in all enabled AZs - Zonal DNS Name - NLB has DNS names for each of its nodes - Use to determine the IP address of each node - Used to minimies latency and data transfer costs - You need to implement app specific logic

Gateway Load Balancer - Deploy, scale and manage a fleet of third party network virtual appliances in AS - Example: Firewalls, Intrusion Detection and Prevention Systems, Deep Packet Inspection Systems, payload manipulation.... - Operates at Layer 3 (Network Layer) - IP Packets - Combines the following functions - Transparent Network Gateway - single entry/exit for all traffic - Load Balancer - distributes traffic to you virtual appliances - Uses the GENEVE protocol on port 6081

Gateway Load Balancer - Target Groups - EC2 Instances - IP Addresses - must be private IPs

4.15 Elastic Load Balancers - Part 2

Cross-Zone Load Balancing

- With cross zone load balancing - each load balancer instance distributes evenly across all registered instances in all AZ - Without Cross Zone Load Balancing - Requests are distributed in the instances of the node of the Elastic Load Balancer

- Classic Load Balancer - Disabled by default - No charges for inter AZ data if enabled - Application Load Balancer - Always on (can’t be disabled) - No charges for inter AZ data - Network

Load Balancer - Disabled by default - You pay charges for inter AZ data if enabled - Gateway Load Balancer - Disabled by default - You pay charges for inter AZ data if enabled

Sticky Sessions (Session Affinity) - It is possible to implement stickiness so that the same client is always redirected to the same instance behind a load balancer - This works for Classic Load Balancers and Application Load Balancers - The ”cookie” used for stickiness has an expiration data you control - Use case: make sure the user doesn’t lose his session data - Enabled stickiness may bring imbalance to the load over the backend EC2 instances.

Request Routing Algorithms - Least Outstanding Requests - The next instances to receive the requst is the instance that has the lowest number of pending / unfinished requests - Works with Application Load Balancer and Classic Load Balancer (HTTP / HTTPS)

Request Routing Algorithms - Round Robin - Equally choose the targets from the target group - Works with Application Load Balncer and Class Load Balancer

Request Routing Algorithsm - Flow Hash - Selects a target based on the protocol, source / destination IP address source/ destination port and TCP sequence number - Each TCP/ UDP connection is routed to a single target for the life of the connection - Works with Network Load Balancer

4.16 API Gateway

API Gateway - Overview - Helps expose Lambda, HTTP and AWS Services as an API - API versioning, authorisation, traffic management (API Keys, throttles), huge scale, serverless, req/resp transformations, OpenAPI spec, CORS - Limits to know - 29 Seconds timeout - 10 MB max payload size

API Gateway - Deployment Stages - API changes are deployed to ”stages” (as many as you want) - Use the naming you like for stages (dev, test, prod) - Stages can be rolled back as a history of deployments is kept

API Gateway - Integrations - HTTP - Expose HTTP endpoints in the backend - Example: internal HTTP API on premise, Application Load Balancer - Why? Add rate limiting, caching, user authentication, API keys, etc..... - Lambda Function - Invoke Lambda function - Easy way to expose REST API backed by AWS lambda - AWS Service - Expose any AWS API through the API Gateway? - Example: Start an AWS Step Function workflow, post a message to SQS - Why? Add authentication, deploy publicly. rate control.......

Solution Architecture Discussion: API Gateway in front of S3

API Gateway - Endpoint Types - Edge-Optimised (default): For gobal clients - Reqeusts are routed through the CloudFront Edge Locations (improves latency) - The API Gateway still lives in only one region - Regional: - For clients within the same region - Could manually combine with CloudFront (more control over the caching strategies and the distribution) - Private: - Can only be access from your VPC using an interface VPC endpoint (ENI) - Use a resource policy to define access

Caching API responses - Caching reduces the number of calls made to the backend - Default TTL is 300 seconds (min 0s, max is 3600s) - Caches are defined per stage - Possible to override cache settings —¿ per method - Clients can invalidate the cache with header: Cache-Control:max-age=0 (with proper IAM authorisation) - Able to flush the entire cache (invalidate it) immediately. - Cache encryption option - Cache capacity between 0.5GB to 237GB

API Gateway - Errors - 4xx means client errors - 400: Bad Requst - 403: Access Denied: WAF filtered - 429: Quote exceeded, Throttle - 5xx means Server errors - 502: Bad Gateway Exception, usually for an incompatible output returned from a Lambda proxy integration backend and occasionally for out-of-order invocations due to heavy loads. - 503: Service Unavailable Exception - 504: Integration Failure - ex Endpoint Request Timed-out Exception —¿ API Gateway requests time out after 29 seconds maximum

API Gateway - Security - Load SSL certificates and use Route53 to defined a CNAME - Resource Policy (S3 Bucket Policy) - control who can access the API - Users from AWS accounts, IP or CIDR blocks, VPC or VPC endpoints - IAM Execution Roles for API gateway at the API level - To invoke a Lambda Function, an AWS services.............. - CORS (Cross-origin resources sharing) - Browser based security - Control which domains can call your API

API Gateway - Authentication - IAM based access (AWS$_IAM) - Goodforprovidingaccesswithinyourinfrastru$
$PassIAMcredentialinheadersthroughSigV4 - LambdaAuthoriser(formerlyCustomerAuthoriser) -$
$UseLambdatoverfifyacustomOAuth/SAML/thridpartyauthentication - Cognitouserpools - Clientauthenticate$
$ClientpassesthetokentotheAPIGateway - APIGatewayknowout-of-the-boxhowtoverifytotoken$

API Gateway - Logging, Monitoring, Tracing - CloudWatch logs: - Enabled CloudWatch logging at the Stage level (with Log Level - ERROR, INFO) - Can log full reqeusts / responses data - Can send API Gateway Access Logs (customisable) - Can send logs directly into Kinesis Data Firehose (as an alternative to CW logs) - CloudWatch Metrics - Metrics are by stage, possibility to enable detailed metrics - IntegrationLatency, Latency, CacheHitCount, CacheMissCount - X-Ray: - Enable tracing to get extra information about requests min API Gateway - X-Ray API Gateway + AWS Lambda give you the full picture.

4.17 API Gateway - Part 2

AWS Gateway - Usage Plans and API Keys - If you wan to make an API available as an offering to your customers - Usage PLan: - Who can access on or more deployed API stages and methods - how much and how fast they can access them - uses API keys to identify API clients and meter accesss - configure throttling limits and quota limits that are enforced on individual clients - API Keys: - alphanumeric string values to distribute to you customers - API key here...... - Can use with usage plans to control access - Throtling limits are applied to the API keys - Quotes limits is the overall number of maximum reqeusts - 429 Too Many Requests - Acount level throttling across all APIs in a region - Clients must implement retry mechanisms

API Gateway - WebSocket API - Overview - What is a WebSocket? - Two-way interactive communication between a uses's browser and a server - Server can push information to the client - This enables stateful application use cases - WebSocket APIs are often used in real-time applications such as chat applications, collaboration platforms, multiplayer games and financial trading platforms. - Works with AWS services (Lambda, DynamoDB) or HTTP endpoints.

Server to Client Messaging - @conenctions used for replies to clients

API Gateway - Private APIs - Can only be accessed from your VPC by using a VPC Interface Endpoint - Each VPC Interface Endpoint can be used to access multiple Private APIs

- API Gateway Resource Policy - Allow or deny access to API from selected VPCs and VPC Endpoints, including across AWS accounts - aws:SourceVpc and aws:SourceVpce

4.18 AWS AppSync

- AppSync is a managed service that uses GraphQL - GraphQL makes it easy for applications to get exactly the data they need. - This includes combining data from one or more sources - NoSQL data stores, Relational databases, HTTP APIs....... - Integrates with DynamoDB, Aurora, ElasticSeach and others - Customer sources with AWS Lambda - Retrieve data in real-time with WebSocket or MQTT or WebSocket - For mobile apps: local data access and data synchronisation - It all start with uploading a GraphQL Schema

AppSync Diagram

AppSync - Cognito Integration - Perform authorisation on Cognito users based on the groups they belong to. - In the GraphQL schema, you can specify the security for Cognito groups

4.19 Route 53 - Part 1

Route 53 - Record Types - A - maps a hostname to IPv4 - AAAA - maps a hostname to IPv6 - CNAME - maps a hostname to another hostname - The target is a domain name which must has an A or AAAA record - Can't create a CNAME record for the top node of a DNS namespace (Zone Apex) - Example: can't create example.com but you can create www.example.com - NS - Name Servers for the Hosted Zone - Control how traffic is routed for a domain

Route 53 - CNAME vs. Alias - AWS Resources (Load Balancer, CloudFront...) expose an AWS hostname: - CNAME - Points a hostname to any other hostname —–¿ Only for non root domain!!! - Alias - Points a hostname to an AWS resource —-¿ Works for root domain and non root domain - Free of charge - Native health check

Route 53 - Alias Records Targets - Elastic Load Balancers - CloudFront Distributions - API Gateway - Elastic Beanstalk environments - S3 Websites - VPC Interface Endpoitns - Global Accelerator accelerator - Route 53 record in the same hosted zone —-¿ You cannot set an ALIAS record for an EC2 DNS name

Route 53 - Records TTL (Time To Live) - High TTL - e.g 24hr - Less traffic on Route 53 - Possibly outdated records - Low TTL - e.g 60 seconds - More traffic on Route 53 –¿ more cost - Records are outdated for less time - Easy to change records - Except for alias records TTL is mandatory for each DNS record

Routing Policies - Simple - Typically route traffic to a single resource - Can't be associated with Health Checks - Can specify multiple values in the same record - If multiple values are returned, a random one is chose by the client.

Routing Policies - Weighted - Control the % of the requests that go to each specific resource - Can be associated with health checks - Use cases: load balancing between regions, testing new application versions.

Routing Policies - Latency based - Redirect the resource that has the least latency close to us - Super helpful when latency for users is a priority - Latency is based on traffic betwen users and AWS regions — Germany users may be directed to the US (if that is the lowest latency) — Can be associated with Health Checks (has failover capbility)

Routing Policies - Failover (Active-Passive)

Routing Policies - Geolocation - Different from Latency-based - This routing is based on user location - Specify location by Continent, Country or by US State (if theres overlapping, most precise location selected) - Should create a "default" record (in case theres no match on location) - Use cases: website localisation, restrict content distribution, load balancing...... - Can be associate with Health Checks

Routing Policies - Geoproximity - Route traffic to your resources based on the geographic location of users and resources - Ability to shift more traffic to resources based on the defined bias - To change the size of the geographic region specify bias values: - To expand (1 to 99) - more traffic to the resource - To shrink (-1 to -99) - less traffic to the resource - Resources can be:

- AWS resources (specify AWS region) - Non-AWS resources (specify Lattitude and Longitude) - You must use Route 53 Traffic Flow to use this feature

Route 53 - Traffic Flow - Simplify the process of creating and maintaining records in large an complex configurations - Visual editor to manage complex routing decision trees - Configurations can be saved as Traffic Flow Policy - Can be applied to different Route 53 Hosted Zones (different domain names) - Support versioning

Routing Policies - Multi-Value - Use when routing traffic to multiple resources - Route 53 return multiple values/resources - Can be associated with Health Checks (return only values for healthy resources) - Up to 8 healthy records are returned for each Multi-Value query - Multi-value is not a substitute for having a ELB

Routing Policies - IP- Based Routing - Routing is based on clients IP addresses - You provide a list of CIDRs for you clients and the corresponding endpoints/locations (user-IP-to-endpoing-mappings) - Use cases: Optimise performance, reduce network costs...... - Example: route end users from a particular ISP to a specific endpoint

4.20 Route 53 - Part 2

Route 53 - Hosted Zones - A container for records that define how to route traffic to a domain and its subdomain - Public Hosted Zones - contains records that specify how to route traffic on the Internet (public domain names) - Private Hosted Zone - contains records that specify how you route traffic within one or more VPCs (private domain names)

Route 53 - Public vs. Private Hosted Zones

Route 53 - Good to Know - For internal private DNS (Private Hosted Zone), you must enable the VPC settings enableDNSHostnames and enableDNSSupport - DNS Security Extensions (DNSSEC) - A protocol for securing DNS traffic, verifies DNS data integrity and origin - Protects agains Man in the Middle (MITM) attacks - Route 53 supports both DNSSEC for Domain Registration and DNSSEC Signing - Works only with Public Hosted Zones - Route 53 with third registrar - you can buy the domain out of AWS and use Route 53 as the DNS provider - Update the NS records on the thrid party Registrar

Route 53 - Health Checks - HTTP Health Checks are only for public resources - Health Check -¿ Automated DNS Failover: - Health checks that monitor an endpoint (application, server, other AWS resource) - Health checks that monitor other health checks (Calculated Health Checks) - Health checks that monitor CloudWatch Alarms – e.g throttles DynamoDB, alarms on RDS, customer metrics (useful for private resources) - Health Checks are integrated with CW metrics

Route 53 - Calculated Health Checks - Combine the results of multiple Health Checks into a single health check - You can use OR, AND or NOT - Can monitor up to 256 Child Health Checks - Specify how many of the health checks need to pass to make the parent pass. - Usage: perform maintenance to you website withouth causing all health checks to fail.

Health Checks - Monitor an Endpoint - About 15 global health checkers wiill check the endpoint health - Health Checks pass only when the endpoint responds with the 2xx and 3xx status codes. - Health Checks can be setup to pass / fail based on the text in the first 5120 bytes of the response.

Health Checks - Private hosted Zones - Route 53 health checkers are outside the VPC - They can't access private endpoints (private VPC or on-premises resource) - You can create a CloudWatch Metric and associate a CloudWatch Alarm then create a Health Check that checks the alarm itself.

Health Checks Solutions Architecture - RDS multi-region fail over

4.21 Route 53 - Resolvers and Hybrid DNS

- By default, Route 53 Resolver automatically answers DNS queries for: - Local domain names for EC2 instances - Records in Private Hosted Zones - Records in public Name Servers - Hybrid DNS - resolving DNS queries between VPC (Route 53 Resolver) and your networks (other DNS Resolvers) - Networks can be: - VPC itself / Peered VPC - On-premises Network (connected through Direct Connect or AWS VPN)

Route 53 - Resolver Endpoints - Inbound Endpoint - DNS Resolvers on your network can forward DNS queries to Route 53 Resolver - Allows you DNS Resovlers to resolve domain names for AWS resources (e.g. EC2 instances) and records in Route 53 Private Hosted Zones - Outbound Endpoint - Route 53 Resolver conditionally forwards DNS queries to you DNS Resolvers - Use Resolver Rules to forward DNS queries to you DNS Resolvers - Associated with one or more VPC's in the same AWS Region - Create in two AZs for high availability - Each Endpoint support 10,000 queries per second per IP address

Route 53 - Resolver Inbound Endpoints

Route 53 - Resolver Outbound Endpoints

Route 53 - Resolver Rules - Control which DNS queries are forwarded to DNS Resolvers on your network - Conditional Forwarding Rules (Forwarding Rules) - Forward DNS queries for specified domain and all its subdomains to target IP addresses - System Rules - Selectively overriding the behaviour defined in Forwarding Rules (e.g don't forward DNS queries for a subdomain acme.example.com) - Auto-Defined System Rules - Defines how DNS queries for selected domains are resovled (e.g AWS internal domain names, Private Hosted Zones) - If multiple rules matched, Route 53 Resolver chooses the most specific match. - Resolver Rules can be shared across accounts using AWS ”RAM” - Manage them centrally in one account - Send DNS queries from multiple VPC to the target IP defined in the rule

4.22 AWS Global Accelerator

4.23 Comparison of Solution Architecture

4.24 AWS Outposts

4.25 AWS WaveLength

4.26 AWS Local Zoness

Chapter 5

Storage

5.1 EBS and Local Instance Store

5.2 Amazon EFS

5.3 Amazon S3

5.4 Amazon S3 - Storage Class Analysis

5.5 Amazon S3 - Storage Lens

5.6 S3 Solution Architecture

5.7 Amazon FSx

5.8 Amazon FSx - Solution Architecture

5.9 AWS DataSync

5.10 AWS DataSync - Solution Architecture

5.11 AWS Data Exchange

5.12 AWS Transfer Family

5.13 AWS Storage Services Price Comparison

Chapter 6

Caching

6.1 Cloudfront - Part 1

- Content Delivery Network (CDN) - Improves read performance, content is cached at the edge - 225+ Point of Presensce globally (215+ Edge Locations and 13 Regional Edge Caches) - Protect against Network and Application layer attacks (e.g DDoS attacks) - Integration with AWS Shield, AWS WAF and route 53 - Can expose external HTTPS and can talk to internal HTTPS backends - Supports WebSocket protocol

6.1.1 CloudFront - Origins

- S3 Bucket - For distributing files - For uploading to S3 (using CloudFront as an ingres) - Enhanced security with CloudFront Origin Access Control (OAC) - MediaStore Container and MediaPackage Endpoint - To deliverVideo on Demand (VOD) or live streaming video using AWS Media Services - VPC Origin - For applications hosted in VPC private subnets - Application Load Balance / Network Load Balancer / EC2 Instances - Customer Origin (HTTP) - API Gateway (for more control.... otherwise use API Gateway Edge) - S3 Bucket configured as a website (enable Static Website hosting) - Any HTTP backend you want

6.1.2 CloudFront - S3 as an Origin

6.1.3 CloudFront vs S3 Cross Region Replication

- CloudFront: - Global Edge network - Files are cached for a TTL (maybe a day) - Great for static content that must be available everywhere

- S3 Cross Region Replication - Must be setup for each region you want replication to happen - Files are updated in near real-time - Read only - Great for dynamic content that needs to be available at low-latency in few regions

6.1.4 CloudFront - ALB or EC2 as an origin Using VPC Origins

- Allows you to deliver content from you applications hosted in your VPC private subnets (no need to expose them on the Internet) - Deliver traffic to private - Application Load Balancer - Network Load Balancer - EC2 Instances

6.1.5 CloudFront - EC2 or ALB as an origin

6.1.6 CloudFront - Restrict Access to Application Load Balancers and Custom Origins

- Prevent direct access to you ALB or Custom Origins (only access through CloudFront) - First, configurations CloudFront to add a CustomHTTPHeader to requests it sends to the ALB - Second, configure the ALB to only forward requests that contain that Customer HTTP Header - Keep the custom header name and value secret!

6.1.7 CloudFront - Origin Groups

- To increase high-availability and do failover - Origin Group: one primary and on secondary - If the primary origin fail; the second on is used - Origins can be cross AWS regions

6.2 Cloudfront - Part 2

6.2.1 CloudFront Geo Restrictions

- You can restrict who can access your distribution - Allow list: Allow you users to access your content only if they're in one of the countries on a list of approved - Block list: Prevent your users from accessing your content if they're in one of the countries on a blacklist of banned countries..... lol - The ”country” is determined using a third party Geo-IP database - Use case: Copyright Laws to control access to content - Note: the geo header CloudFront-Viewer-Country is in Lambda at Edge

6.2.2 CloudFront - Pricing

- CloudFront Edge locations are all around the world - The cost of data out per edge location varies

6.2.3 CloudFront - Price Classes

- You can reduce the number of edge locations for cost reductions - Three price classes - Price Class All: all regions - best performance - Price Class 200: most regions, but excludes the most expensive regions - Price Class 100: only the least expensive regions

6.2.4 CloudFront Signed URL Diagram

- Signed URL with expiration to control access to content in CloudFront - The Signed URL are generated by an API call into CloudFront as a trusted signer

6.2.5 CloudFront Signed URL vs S3 Pre-Signed URL

- CloudFront Signed URL: - Allow access to a path, no matter the origin - Account wide key-pair, only the root can manage - Can filter by IP, path, date, expiration - Can leverage caching features

- S3 Pre-Signed URL: - Issue a request as the person who pre-signed URL - Uses the IAM key of the signed IAM principal - Limited lifetime

6.2.6 CloudFront - Custom Error Pages

- Return an object to the viewer (e.g html) when your origin returns an HTTP 4xx or 5xx status code to CloudFront - Use Error Caching Minimum TTL to specify how long CloudFront caches the customer error pages

6.3 Lambda at Edge and CloudFront Functions

6.3.1 CloudFront - Customisation at the edge

- Many modern application execute some form of the logic at the edge - Edge function: - A code that you write an attach to CloudFront distributions - Runs close to your users to minimise latency - Doesn't have any cache, only to change requests/ responses - CloudFront provides two types: CloudFront Functions and Lambda and Edge Use cases: - Manipulate HTTP requests and responses - Implement request filtering before reaching your application - User authentication and authorisation - Generate HTTP resources at the edge - A/B Testing - Bot mitigation - You don't have to manage any servers, deployed globally

6.3.2 CloudFront Functions and Lamabda at Edge

6.3.3 CloudFront - CloudFront Functions

- Lightweight functions written in JavaScript - For high-scale latency-sensitive CDN customisation - Sub-ms startup times, million of requests per second - RuN at Edge Locations - Processed-Based isolation - Used to changeViewer requests and responses: - Viewer Requests: after CloudFront recevives a reqeusts from a viewer - Viewer Response: before CloudFront forwards the response to the viewer - Native feature of CloudFront (manage code entirely within CloudFront)

6.3.4 CloudFront - Lambda at Edge

- Lambda functions written in NodeJS or Python - Scala to 1000s of reqeust/ second - Runs at the nearest Regional Edge Cache - VM-based isolation - Used to change CloudFront requests and responses - Viewer Request - after CloudFront recieves a reqeust from a viewer - Origin Request - before CloudFront forwards the response to the origin - Orign response - after CloudFront recieves the response from the origin - Viewer Response - before CloudFront forwards the response to the viewer - Author your functions in one AWS Region (some region) then CloudFront replicats to its locations

6.3.5 CloudFront Functions with Lambda at Edge

CloudFront Functions and Lambda at Edge can be used together Note: You can't combine CloudFront Functions and Lambda at Edge in viewer events (viewer reqeust and viewer response)

6.3.6 Using Lambda at Edge only

Use when you need some of the capabilites of Lambda at Edge that aren't available with CloudFront Functions (e.g longer execution time, network access,....)

6.3.7 CloudFront Funtions vs. Lamda at Edge

6.3.8 CloudFront Funcitons vs Lambda at Edge Use Cases

- CloudFront Functions - Cache key normalisation - Transform reqeust attributes (headers, cookies, query string, URL) to create an optimal Cache Key - Header manipulation - Insert/modify/ delete HTTP headers in the request or response - URL rewrites or redirects - Request authentication and authorisation - Create and validate user-generated tokens (e.g JWT) to allow/deny reqeusts

Lambda at Edge - Longer execution time (several ms) - Adjustable CPU or memory - You code depends on a third libraries (e.g AWS SDK to access other aws Services) - Network access to use external services for processing - File system access or access to the body of HTTP requests

6.3.9 CloudFront Functions vs. Lambda at Edge - Authentication and Authorisation

CloudFront Functions

Lambda at Edge

6.3.10 Lambda@Edge: Loading content based on User-Agent

6.3.11 Lambda at Edge - Global Application

6.4 Lambda at Edge Reduce Latency

6.4.1 Lambda at Edge - Route to different origin

6.5 Amazon ElastiCache

6.5.1 Amazon ElasticCache Overview

- The same way RDS is to get managed Relational Databases..... - ElastiCache is to get managed Redis or Memcached - Caches are in memory databases with really high performance and low latency - Helps reduce load off of databases for read intensive workloads - Helps make you application stateless - AWS takes care of OS maintenance / patching, optimisations, setup, configuration, monitoring failure recovery and backups. - **Using ElastiCache involves heavy application code changes**

6.5.2 ElastiCache Solution Architecture - DB Cache

- Applications queries ElastiCache, if not available, get from RDS and store in ElasticCache - Helps relive load in RDS - Cache must have an invalidation strategy to make sure only the most current data is used there.

6.5.3 ElastiCache Solution Architecture - User Session Store

- User logs into any of the application - The application writes the session data into ElastiCache - The user hits another instance of our application - The instance retrieves the data the user is already logged in

6.5.4 ElastiCache - Redis vs Memcached

- Redis - Multi AZ with Auto-Failover - Read Replicas to scale reads and have high availability - Persistent, Data Durability: Append Only File (AOF), backup and restore features - Memcached - Multi-node for partitioning of data (sharding) - Non Persistent - Backup and restore (Serverless) - Mult-threaded architecture

6.6 Handling Extreme Rates

Chapter 7

Databases

7.1 DyanamoDB

7.2 Amazon OpóenSearch

7.3 RDS

7.4 Aurora - Part 1

7.5 Aurora - Part 2

Chapter 8

Service Communication

8.1 Step Functions

8.2 SQS

8.3 Amazon MQ

8.4 Amazon SNS

8.5 Amazon SNS - SQS Fan Out Pattern

8.6 Amazon SNS - Message Delivery Retries

Chapter 9

Data Engineering

9.1 Amazon Kinesis Data Streams

9.2 Amazon Data Firehose

9.3 Amazon Managed Service for Apache Flink

9.4 Streaming Architectures

9.5 Amazon MSK

9.6 AWS Batch

9.7 Amazon EMR

9.8 Running Jobs on AWS

9.9 AWS Glue

9.10 Redshift

9.11 Amazon DocumentDB

9.12 Amazon Timestream

9.13 Amazon Athena

9.14 Amazon QuickSight

9.15 Big Data Architecture

Chapter 10

Monitoring

10.1 CloudWatch

10.2 CloudWatch Logs

10.3 Amazon EventBridge

10.4 X-Ray

10.5 AWS Personal Health Dashboard

Chapter 11

Deployment and Instance Management

11.1 Elastic Beanstalk

11.2 CodeDeploy

11.3 CloudFormation

11.4 Service Catalog

11.5 SAM - Serverless Application Model

11.6 AWS CDK - Cloud Development Kit

11.7 AWS Systems Manger - SSM

11.8 AWS Cloud Map

Chapter 12

Cost Control

12.1 Cost Allocation Tags

- With Tags we can track resources that relate to each other
- With Cost Allocation Tags we can enable detailed costing reports
- Just like Tags, but they show up as columns in Reports
- AWS Generated Cost Allocation Tags
  - Automatically applied to the resource you create
  - Start with Prefix aws: (e.g. aws: createdBy)
  - They're not applied to resources created before the activation
- User tags
  - Defined by the user
  - Start with Prefix user:
- Cost Allocation Tags just appear in the Billing Console
- Takes up to 24 hours for the tags to show up in the report

12.2 AWS Tag Editor

- Allows you to managed tags of multiple resources at once - You can add/update/delete tags - Search tagged/untagged resources in all AWS Regions

12.3 Trusted Advisor

- No need to install anything - high level AWS account assessment - Analyse your AWS accounts and provides recommendation - Cost Optimisation - Performance - Security - Fault Tolerance - Service Limits - Operational Excellence - Core Checks and recommendations - all customers - Can enable weekly email notification from the console - Full Trusted Advisor - Available for Business and Enterprise support plans - Ability to set CloudWatch alarms when reaching limits - Programmatic Access using AWS support API

Column1	Basic Support
AWS Trusted Advisor Best Practice Checks	7 Core Checks
Enhanced Technical Support	24x7 customer service, documentation, whitepapers and support
Case Severity / Response Times	Data12
Data16	Data17

Table 12.1: Example Table with 4 Rows and 5 Columns

Trusted Advisor - Good to Know

- Can check if an S3 bucket is made public - But cannot check for S3 objects that are public inside of your bucket - Use Amazon EventBridge / S3 Events instead / AWS Config Rules

- Service Limits - Limits can only be monitored in Trusted Advisor (cannot be changed) - Cases must be created manually in AWS Support Centre to increase limits - OR use the AWS Service quotas service

12.4 AWS Service Quotas

- Notify you when you’re close to a service quota value threshold - Create CloudWatch Alarms on the Service Quotas console - Example: Lambda concurrent executions - Helps you know if you need to request a quota increase or shutdown resources before limit is reached

12.5 EC2 Launch Types and Savings Plans

- On Demand Instances - short workload, predictable pricing, reliable. - Spot Instances - short workloads for check, can lose instances (not reliable) - Reserved: (Minimum 1 year) - Reserved Instances - long workloads - Convertible Reserved Instances - long workloads with flexible instances - Dedicated Instances: no other customers will share you hardware - Dedicated Hosts: book an entire physical server, control instance placement - Great for software licenses that operate at the core, or socket level - Can define host affinity so that instance reboots are kept on the same host

12.5.1 AWS Savings Plan

- New pricing model to get a discount based on long-term usage - Commit to a certain type of usage: ex $10 per hour for 1 to 3 years - Any usage beyond the savings plan is billed at the on-demand price

- EC2 Instance Savings plan (72% - same discount as Standard RIs) - Select instance family and locked to a specific region - Flexible across size, OS (Windows to Linux) tenancy. (dedicated or default) - Compute Savings Plan - Ability to move between instance family, region, compute type and OS and tenancy - SageMaker Savings plan (up to 64% off)

12.6 S3 Cost Savings

12.6.1 S3 Storage Classes

- Amazon S3 Standard - General Purpose - Amazon S3 Standard-Infrequent Access (IA) - Amazon S3 One Zone-Infrequent Access - Amazon S3 Glacier Instant Retrieval - Amazon S3 Glacier Flexible Retrieval - Amazon S3 Glacier Deep Archive - Amazon S3 Intelligent Tiering

Can move between classes manually or using S3 lifecycle configurations

12.6.2 S3 - Other Cost Savings

- S3 Lifecycle Rules: transition objects between tiers - Compress Objects - to save space - S3 Requester Pays: - In general, bucket owners pay for all Amazon S3 storage and data transfer costs associated with their bucket - With Requester Pays buckets, the requester instead of the bucket owner pays the cost of the request and the data downloaded from the bucket - The bucket owner always pays the cost of storing data - Helpful when you want to share large datasets with other accounts - If an IAM role is assumed the owner account of that role pays for the request

12.7 S3 Storage Classes - Reminder

12.7.1 S3 Storage Classes

- Amazon S3 Standard - General Purpose - Amazon S3 Standard-Infrequent Access (IA) - Amazon S3 One Zone-Infrequent Access - Amazon S3 Glacier Instant Retrieval - Amazon S3 Glacier Flexible Retrieval - Amazon S3 Glacier Deep Archive - Amazon S3 Intelligent Tiering

Can move between classes manually or using S3 lifecycle configurations

12.7.2 S3 Durability and Availalbility

- Durability: - High Durability (99.9999999999, 11 nines) of objects across multiple AZ - If you store 10,000,000 object with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000 years (nice....but basic math) - Same for all storage classes

- Availability - Measures how readily available a service is - Varies depending on storage class - Example: S3 standard has 99.99% availability = not available 53 minutes a year

S3 Standard - General Purpose - 99.99% Availability - Used for frequently accessed data - Low latency and high throughput - Sustain 2 concurrent facility faiures - Use Cases: Big Data analytics, mobile and gaming applications, content and distribution

S3 Storage Classes - Infrequent Access - For data that is less frequently accessed, but requires rapid access when needed - Lower cost than S3 standard

Amazon S3 Standard-Infrequent Access (S3 Standard-IA) - 99.9% Availability - Use cases: Disaster Recovery, backups

Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA) - High durability (99.999999999) in a single AZ; data lost when AZ is destroyed - 99.5% Availability - Use Cases: Storing secondary backup copies of on-premise data or data you can recreate

Amazon S3 Glacier Storage Classes - Low-cost object storage meant for archiving / backup - Pricing: price for storage + object retrieval cost

- Amazon S3 Glacier Instant Retrieval - Millisecond retrieval, great for data accessed once a quarter - Minimum storage duration of 90 days - Amazon S3 Glacier Flexible Retrieval (Formerly

Amazon S3 Glacier) - Expedited (1 to 5 minutes), Standard (3 to 5 hours), Bulk (5 to 12 hours) - free - Minimum storage duration of 90 days - Amazon S3 Glacier Deep Archive - for long term storage - Standard (12 hours), Bulk (48 Hours) - Minimum Storage duration of 180 days

12.7.3 S3 Intelligent Tiering

- Small monthly monitoring and auto-tiering fee - Moves objects automatically between Access Tiers based on usage - There are no retrieval charges in S3 Intelligent-Tiering

- Frequent Access tier (automatic): default tier - Infrequent Access Tier (automatic): objects not accessed for 30 days - Archive Instant Access tier (automatic): objects not accessed for 90 days - Archive Access tier (optional): configurable from 90 days to 700+ days - Deep Archive Access tier (optional): config from 180 days to 700+ days

S3 Storage Classes Comparision

S3 Storage Classes - Price Comparison

12.8 AWS Budgets and Cost Explorer

12.8.1 AWS Budgets

- Create budget and send alarms when costs exceeds the budget - 4 Types of budgets: Usage, Cost, Reservation, Savings Plans - For Reserved Instances (RI) - Track utilisation - Supports EC2, ElastiCache, RDS, Redshift - Up to 5 SNS notifications per budget - Can filter by: Service, Linkedin Account, Tag, Purchase Option, Instance Type, Region, Availability Zone, API Operations, etc...... - Same options as AWS Cost Explorer - 2 budgets are free than $0.002 / day / budget

12.8.2 Budget Actions

- Run actions on your behalf when a budget exceeds a certain cost or usage threshold - Supports 3 actions types - Applying an IAM Policy to a user, group or IAM role - Applying Service Control Policy (SCP) to an OU - Stop EC2 or RDS Instances - Actions can be executed automatically or require a workflow approval process - Reduced unintentional overspending in your account.

12.8.3 Centralised Budget Management

12.8.4 DeCentralised Budget Management

12.8.5 Cost Explorer

- Visualise, understand and manage your AWS costs and usage over time - Create custom reports that analyse cost and usage data - Analyse you data at a high level: total costs and usage across all accounts - Or Monthly, hourly, resource level granularity - Choose an optimal Savings PLan (to lower prices on your bill) - Forcast usage up to 12 months based on previous usage

12.9 AWS Compute Optimiser

- Reduce costs and improve performance by recommending optimal AWS resources for you workloads - Helps you choose optimal configurations and right-size your workloads (over/under provisioned) - Uses Machine Learning to analyse your resources configurations and their utilisation CloudWatch metrics - Supported resources - EC2 Instances - EC2 Auto Scaling Groups - EBS volumes - Lambda functions - Lower your costs by up to 25% - Recommendations can be exported to S3

12.9.1 Computer Optimiser - CloudWatch Agent

- Needed to analyse Memory Utilisation - Not needed for CPU, NetworkIn/Out, DiskReadOps, DiskWriteOps

12.10 EC2 Reserved Instance

- Reserved Instances in an AWS Organisation - All accounts share the Reserved Instances and Savings Plan - The payer account (Management account) of an organisation can turn off Reserved Instance (RI) discount and Savings Plans discount sharing for any accounts in that organisation, including the payer account. - Renewal of Reserved Instances - You can queue (schedule or reserve ahead of time) you reserved instances - To renew a RI, just queue an RI purchase whenever the previous one expires

Chapter 13

Migration

Chapter 14

VPC

14.1 VPC - Basiscs

14.2 VPC - Peering

14.3 Transit Gateway

14.4 VPC Endpoints

ó

14.5 VPC Enpoint Policies

14.6 PrivateLink

14.7 AWS S2S VPN

14.8 AWS Client VPN

14.9 Direct Connect

14.10 On-Premise Redundant Connections

14.11 VPC Flow Logs

14.12 AWS Network Firewall

Chapter 15

Machine Learning

15.1 Rekognition Overview

- Find objects, people, text, scenes in images and videos using ML - Facial analysis and facial search to do user verification, people counting. - Create a database of ”familiar faces” or compare against celebrities - Use cases: - Labeling - Content Moderation - Text Detection - Face Detection and Analysis (gender, age range, emotions....) - Face Search and Verification - Celebrity Recognition - Pathing (ex: for sports game analysis)

15.1.1 Amazon Rekognition - Content Moderation

- Detect content that is inappropriate, unwanted or offensive (image and videos) - Used in social media, broadcast media, advertising and e-commerce situation to create a safer user experience - Set a Minimum Confidence Threshold for items that will be flagged - Flag sensitive content for manual review in Amazon Augmented AI (A2I) - Help comply with regulations

15.2 Transcribe Overview

- Automatically convert speech to text. - Uses a deep learning process called automatic speech recognition (ASR) to convert speech to text quickly and accurately. - Automatically remove Personally Identifiable Information (PII) using Redaction. - Supports Automatic Language Identification for multilingual audio - Use cases: - transcribe customer service calls - automate closed captioning and subtitling - generate metadata for media assets to create a fully searchable archive

15.3 Polly Overview

- Turn text into lifelike speech using deep learning - Allowing you to create applications that talk

15.3.1 Amazon Polly - Lexicon & SSML

- Customise the pronunciation of words with Pronunciation lexicons - Stylized words: St3ph4ne =¿ ”Stephane” - Acronyms: AWS =¿ ”Amazon Web Services” - Upload the lexicons and use them in the SynthesizeSpeech operation. - Generate speech from plain text or from documents marked up with Speech Synthesis Markup Language (SSML) - enables more customisation -

Emphasising specific words or phrases. - Using phonetic pronunciation. - Including breathing sounds, whispering. - Using the Newscaster speaking style

15.4 Translate Overview

- Natural and accurate language translation - Allows you to localise content - such as websites and applications - for international users and to easily translate large volumes of text efficiently.

15.5 Lex + Connect Overview

- Amazon Lex: (same technology that powers Alexa) - Automatic Speech Recognition (ASR) to convert speech to text - Natural Language Understanding to recognise the intent of text, callers - Helps build chatbots and call centre bots - Amazon Connect: - Receive calls, create contact flows, cloud-based virtual contact centre - Can integrate with other CRM systems of AWS - No upfront payments, 80% cheaper than traditional contact center solutions *Hmmmm......*

15.6 Comprehend Overview

- For Natural Language Processing - NLP - Fully managed and serverless service - Uses machine learning to find insights and relationships in text - Language of the text - Extracts key phrases, places, people, brands or events - Understands how positive or negative the text is - Analyses text using tokenization and parts of speech - Automatically organises a collection of text files by topic - Sample use cases: - Analyse customer interactions (emails) to find what leads to a positive or negative experience - Create and groups articles by topics that Comprehend will uncover

15.7 Comprehend Medical Overview

- Amazon Comprehend Medical detects and returns useful information in unstructured clinical text: - Physician's notes - Discharge summaries - Test results - Case notes - Uses NLP to detect Protected Health Information (PHI) - DetectPHI API - Store your documents in Amazon S3, analyse real-time data with Kinesis Data Firehose or use Amazon Transcribe to transcribe patient narratives into text that can be analysed by Amazon Comprehend Medical.

15.8 SageMaker Overview

- Fully managed service for developers / data scientists to build ML models - Typically difficult to do all processes in one place + provision servers

15.9 Kendra Overview

- Fully managed document search service powered by Machine Learning - Extract answers from within a document (text, pdf, HTML, PowerPoint, MS Word, FAQs) - Natural language search capabilities - Learn from user interactions / feedback to promote preferred results (Incremental Learning) - Ability to manually fine-tune search results (importance of data, freshness, customer)

15.10 Personalise Overview

- Fully managed ML-service to build apps with real-time personalised recommendations - Example: personalised product recommendations/re-ranking, customised direct marketing Example: User bought gardening tools, provide recommendations on the next one to buy. - Same technology used by Amazon.com - Integrates into existing websites, applications, SMS, email marketing systems......... - Implement in days, not months (don't need to build, train and deploy ML solutions) - Use cases; retail stores, media and entertainment

15.11 Textract Overview

- Automatically extracts text, handwriting and data from any scanned documents using AI and ML - Extract data from forms and tables - Read and process any type of document (PDFs, images) - Use cases: - Financial Services (e.g, invoices, financial reports) - Healthcare (e.g medical records, insurance claims) - Public Sector (e.g tax forms, ID documents, passports)

15.12 Machine Learning Summary

- Rekognition - face detection, labeling, celebrity recognition
- Transcribe - audio to text (ex: subtitles)
- Polly - text to audio
- Translate - translations
- Lex - build conversational bots - chatbots
- Comprehend - natural language processing
- SageMaker - Machine learning for every developer and data scientist
- Kendra - ML-powered search engine
- Personalise - real-time personalised recommendations
- Textract - detect text and data in documents

Chapter 16

Other Services

16.1 Other Services

16.2 IMPORTANT: CodeCommit Discontinuation

16.3 CI/CD

16.4 Amazon CodeGuru

16.5 Alexa for Business, Lex and Connect

16.6 Kinesis Video Streams

16.7 AWS WorkSpaces

16.8 Amazon AppStream 2.0

16.9 AWS Device Farm

16.10 Amazon Macie

16.11 Amazon SES

16.12 Amazon Pinpoint

16.13 EC2 Image Builder

16.14 Other Services Summary

Chapter 17

Exam Preparation