**AWS Data Storage Overview**

this course will cover many of the storage options for aws

challenges - of storage - time - money - space - requests - organization - flexibility

aws cloud can solve this -

services from aws include - Block storage - File storage - Archives - such as S3 and S3 glacier as well as others

you need the agility to deal with your SLA's -

example tiered data storage allow more flexibility for specific datasets- they are scalable and durable

you have diagnostic and monitoring tools such as aws cloud watch aws cloud trail and aws trusted advisor

compute - storage and troubleshooting services and all combined with AWS

advanced storage features can be added quickly at little extra cost such as tagging for lifecycle permissions

or separation of active data - archived data - IA data

AWS provides transparency between architecture and tools being used - as well as all its device interactions with your data

Resource provisioning on aws is fast - you can create storage instances easily in the console and - sdk and console support allows for you to automate this.

In addition - EFS file systems combined with EBS volumes make it easy to manage your data in the cloud

trade capital for powerful cloud architecture full scaling that you could never do on your own

no guesses on capacity - with cloud computing these problems go away

new resources are a click away and your team can use them in minutes

and deployment can be provided easily as well around the world in any region

**peak application loads -**

in an onsite data center you need to over provision AWS stops the need for this

add EC2's add need attach EBS storage instances as needed to those EC2's and pay only for what you use

Cap optional - severs, buildings ect

Op expenditure - heat, light, power, personnel

Infrastructure as a service greatly reduces these costs

cloud agility also means that you have continued tools upgrades happening by amazon that you can get with one click

you only lose one thing with AWS and that is of course security - as AWS is now involved with your data

**Main storage services -**

S3 - S3 glacier - object storage and archiving

Amazon EBS provides block storage

Amazon EFS - Amazon FSx - provide file storage

Data migration services are provided by the AWS snow family

AWS Snowball - AWS Snowball Edge - AWS Snowmobile

Hybrid storage for on client access is provided by AWS Storage Gateway

and backup is provided by - AWS backup service

**Pricing options -**

in AWS tiered access discounts means often the more you use the less you pay -

this includes services to manage data in addition to the data storage itself

**hybrid cloud -**

this allows a data center to interact with AWS -

aws provides the gateway services and the data center can connect over the internet to AWS with things like API's or VPN's

or AWS storage gateway

**Types of storage -**

**Block storage** - a sever interacting with different volumes of storage

**File storage** - imagine files in a tree structure system - systems like NFS and SMB control access

**Object storage** - is a flat structure where data is presented as objects and exist in a bucket each object has a unique id or key

when an object is changed a whole new object is created and versioning is used to track the changes across all objects

they are accessed through API's can contain metadata and there is no way to perform partial modification on them

**objects are used in S3-**

for block storage - EBS elastic block storage

for file you have - EFS - elastic file storage

**Block Storage -**

in block storage data is broken into blocks -

a block is defined as a set of bytes and is managed as a set of storage

they all have a unique address and do not contain any meta data

each volume behaves like a hard drive -

they are managed by an admin

they are attached to a specific server- and are managed by fiber channel or others

block level storage supports individual formatting such as NTFS for windows

block based storage is typically as direct attached storage and can be used as boot volumes

block storage is used to storage files for things like db and the blocks can be retrieved individually by systems

Block storage when used correctly is typically the fastest type of storage to interact with

types of block storage -

DAS - direct attached storage - server accessing data volumes -exp hard drive

SAN- Storage area network - computer network that allows access to data storage devices - these devices are locked out from the internet

RAID - redundant array of independent discs - is a method for storing data in different places on hard discs -

with raid multiple independent discs are combined for performance- AWS supports raid configurations for amazon EBS

RAID 0 can - stride multiple volumes together - NOTE RAID5 and RAID6 are not recommended for use with AWS services

think of them as collection links to arrays of block storage blocks

Disadvantages-

one storage at a time

no metadata

additional information handled at the application level

you pay for all the blocks weather you use them or not

and if the server attached to your blocks goes down you lose access completely until a new one is attached

Pro use cases -

relational databases

Enterprise applications

president local storage

data warehousing

big data processing

backup and recovery

AWS block storage services -

Amazon EBS - provides persistent block storage -

replication - low latency performance- easy to scale

EC2 instance connected to EBS storage systems

Persistent block storage with your EC2-

4 different volume types -

1 SSD backed volumes - lowest latency - has sub type 1 and 2

best for things like relational databases and noSQL databases

2 HDD backed volumes - highest throughput - has sub type 1 and 2

these are best for applications using steaming IO -

such as data warehousing and log processing

multiple volumes can be connected to a single ec2 instance

plus cloud watch can provide alerts when you need to scale

an EC2 instance store provides temporary storage for your instance -

good for back up or cache data –

**File Storage -**

file folders in a tree structured system is managed by the operator system shared access can be provide by certain systems - like NFS and SMB

data is stored as a single file in a system - in an inverted tree structure- the files can be docs or media

once storage is allocated - the system is mounted and the subdirectories are crated as needed / path names are used to access files

the directories can contain executable files as well

file storage is used by all the common file systems like NTFS - for windows

- it is the widely used data storage system for all network attached storage

challenges are fine when files are small- but when they get large these can be a big problem

- related data access

- data redundancy

- data format dependences

- data inconsistences

examples-

home directories

web serving

application testing environments

media

big data analytics

Container storage -

file storage from AWS

Amazon EFS - provides fully managed shared file system -

to control access you can encrypt data and set access keys

use the console to create a file system - and the attach it to your EC2 instance -

2 perform modes - general and max- max is good for high throughput

amazon Fsx

easy to use simple interface for creating file storage instances - attach then to ec2 instances-

it is windows based -

full support for NTFS -

uses SSD storage for low latency and high compatibility

fully managed windows file servers

automatically encrypted - it takes daily backups

supports DFS namespaces as well -

Amazon Fsx for Luster provides fully managed Luster file system

great for big data and machine learning

a parallel distributed file system -

a system that provides high throughput for handling even peda bytes of data

this is why it is good for high performance computing like ml

**Object based Storage**

objects can different types of data such as files or media -

flat storage where the data is presented as objects

objects can have built in meta data

each object has a built in identifier

data in object storage can be accessed by API or Http calls

this grantees the object will not be lost as it is backed up

in contrast to block and file where data can be changed - object data is immutable -

this makes object data storage great for massive quantities of static data

objects are updated by creating a new object - this can be version controlled

Use cases-

Static website hosting

document storage

Big Data analytics

Backup and Restore systems

Benefits -

scalability - no limits like file and block

fast data retrieval- keys make getting an object a direct request

analogies to valet parking your data as long as you have your ticket your good

better use of meta data - object storage uses both system and user meta data -

cost efficient - at tera and peda byte scale this is the most cost efficient -

also it is the most durable form of data storage

challenges -

interfaces -

app compatibility -

performance - optimized for scale not performance

Netflix and Airbnb use amazon s3 for static object storage

**AWS object based storage service -**

amazon simple storage service - S3 is the best object storage service in the world today

high level durability - and availability

AWS s3 - sets up multiple regions of backup for your data

you can also send data across regions with ease

amazon MAice uses ML to track and organize your data and detect any issue

administration is also highly tooled out with aws to give you what you need

S3 works with lambda to let you use operations

Analytics can be done without moving the data anywhere

Red shift allows you to perform analytics on the data easily in a s3 bucket

automatic scaling means things will always scale up fast

no other services has more partners with built in tools

S3 is highly durable and highly secured - even in transit -

you can protect data in transit with SSL

and with encryption for users- objects are always redundantly stored

amazon s3 checks and repairs any redundant data -

versioning can be used to stored and retrieve all versions of your objects

thus unintended errors can be handled with ease

cross region replication is how the data is backed up to the highest level -

you must add this feature manually it is not automatic

the marker for this is not replicated when the data is

types of S3-

s3 standard - general purpose

amazon s3 intelligent tearing- for data with unknown or changing access patterns

Amazon s3 slandered IA- for data with unknown or changing access patterns

Amazon s3 i zone IA - for long lived but less frequently accessed data

amazon s3 glacier- long term archive and digital preservation

Amazon s3 glacier deep archive - long term archive and digital preservation

a lifecycle upgraded can be added to make it so data is secured and backed up

on multiple availability zones throughout its life cycle -

s3 glacier - supports lifecycle config as well as costs the lowest of all storage

s3 glacier allows zip files and rar to be used as well

s3 glacier requires an access request this can take hors but you can get an accelerated version that reduces this to minutes for more cost

**Hybrid Cloud Storage**

a hybrid deployment comments cloud and on site resources together -

the most common is for the purpose of extending on premise systems with cloud scalability

it is primarily used as an extension tool

in addition an API is used by the client system to access cloud storage

local access of data is required for - hot cold and dirty data - hot new - cold old - dirty mixed

common hybrid use cases -

disaster recovery -

tiered storage for files and volumes -

many workloads can benefit from hybrid cloud storage -

some ML and other big data centers find that cloud and onsite access for data is essential

benefits -

high scalability

Security - data is more secure with network file walls

Performance - things are faster when things are on site

reliability - cached data and onsite data can be accessed more reliably

cost - cost management is improved a bit due to use of the cloud

Cons

hybrid cloud integration can be complicated

integration changes can also be bad

security has some vulnerabilities as if on premises resources are compromised cloud resources can all be compromised as well

AWS storage gateway services

a services for creating a seamless interaction with onsite services and cloud services -

your applications use standard services to access cloud services that then go through the gateway to things like s3

includes things like auto backup and caching to improve latency

it can work with - S3 - volume - or file systems

you can also store data as virtual tapes that then go to s3

**How to choose the best AWS service**

compliance

security

who needs access -

how do they access it -

client location -

backup - and archive

Aws provides network firewalls

server side encryption

con activity options - like dedicated connections

for web applications - its APIs' and S3 supports this well

it is recommended for reading and writing with multiple users that we all use a single gateway to avoid errors in AWS

AWS backup can be used to back up all files

AWS Data migration and transfer options - /////////////////////////////////////////////

migrating data to the cloud can take a while 1gb per second is 100 days to a petabyte

the snow family of services help move data as quickly as possible

AWS snowball

AWS Snowball Edge

AWS snowmobile

AWS snowball is a peda byte scale data transfer service to help move data within the AWS storage centers like s3 easily and quickly

it is easy cost effective and faster than the internet’s highest rate transfer options

snowball is a service where AWS send you a snowball device you fill it with your data directly and

they ship back to data centers and upload it to your s3 storage directly

AWS snowball edge -

this provides you with special aws data center equipment that go with you track and store all your data on

say a voyage at sea and then be sent back to warehouses latter for full s3 storage

you can transfer - 100petabytes of data with aws snowmobile - a truck data center you can rent from AWS

AWS data sync helps sync up data operations

this optimizes data validation and other features

transfers ten times faster than open source tools

uses NFS protocols

getting started with data sync is as easy as deploying the device in your center and transferring over the data

AWS transfer for SFTP

transfer files on premise to s3

uses SFTP protocol

no SFTP servers to manage

AWS VPN

allows a secure tunnel connection to the aws cloud

AWS direct connect

dedicated network connection to the aws cloud

aws cli and aws sdk's can be used to set up automation requirements

more course on data storage to check out

https://www.aws.training/Details/Video?id=15883