Global Knowledge®

Expert Reference Series of White Papers

AWS Storage Solutions 101

AWS Storage Solutions 101

Chris Littlefield, AWS Certified Solutions Architect – Associate Level, AWS Certified SysOps Administrator – Associate Level

Introduction

In this white paper, we will discuss the variety of storage solutions available from Amazon Web Services (AWS). Developing a broad understanding of the storage capabilities at AWS will enable cloud architects to design architectures that ensure that cloud deployments are highly scalable and available. The goal is to deliver applications in the cloud in the most efficient, cost-effective, and secure manner. Third-party options are also valid within your AWS ecosystem, but for this paper we will limit the focus to the AWS native storage services.

In each section, we introduce a different product AWS provides to store data in the cloud. Then we'll discuss some of the use cases each storage solution is ideal for. Finally we'll give a high-level summary of the security measures you can implement to secure them. For additional service specific information, please follow the links provided at the end of each section.

Block Storage

The Elastic Compute Cloud (EC2) is the virtual server at AWS. EC2 instances support Windows, Linux, and FreeBSD. When we spin up EC2 instances, we store at least the OS data on block storage attached to the EC2 instance. Administrators can choose to use multiple drives for their virtual instance, and the options depend upon the EC2 instance type they select.

There are two types of block storage for EC instances. The first block storage option for EC2 is the instance store, and the second option is the Elastic Block Store (EBS) volume. The root volume can be either instance store, or EBS backed. The secondary volumes can be either instance store or EBS volumes, or a mix of both. Let's look at each EC2 block storage option in turn.

Amazon EC2 Instance Stores:

Instance storage is ephemeral block storage that is directly attached to the EC2 instance. So the instance store and the EC2 instance are running on the same hardware and hypervisor. Instance storage provides administrators with various sized drives based on the instance type. Instance storage is not available with all EC2 instance types. On the EC2 instance types where instance storage is an option, the cost is inclusive of the EC2 instance hour price. You will find standard drive performance and even SSD drives available for your EC2 instances.

When you work with instance stores, you need to take into account the volatile nature of the instance store. The instance store's lifetime is tied directly to the EC2 instance lifetime, so the data is lost when you stop or terminate the associated EC2 instance. Data is also lost if the underlying hardware the instance is running on fails. There are no built-in backup options for instance storage, but we can still backup the data using OS level tools.

By default, the instance store is EXT3 formatted, and can then be formatted with any file system you wish. The key advantage of instance storage over EBS storage is cost, and access to much larger disks with zero network variability.

Security in Brief—EC2 Instance Storage:

You can launch instances in the Virtual Private Cloud (VPC). You can choose to encrypt your instance store volumes. You can use any encryption tools for file systems you wish. You should also consider securing your instance access via SSL/TLS and configuring your security group(s) with a tight, least privilege approach.

Ideal for:

- Temporary storage, stateless web servers, tempdb storage, and as a worker node in a Hadoop cluster.
- For more information see:
  http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/InstanceStorage.html

Amazon Elastic Block Store (EBS)

Amazon's Elastic Block Store volumes provide EC2 instances with block storage that's network attached and persistent. You can format the drive with any file system you wish. EBS volumes can store from one gigabyte up to one terabyte of data per volume. If and when you need more storage you can grow your volume, and/or attach multiple EBS volumes to the same instance. Unlike instance storage, the data on your EBS volumes can persist beyond the life of the EC2 instance. You can ensure your data is safe by detaching the volume prior to terminating the EC2 instance associated with it. The EBS volume is then available to attach to a different EC2 instance for use. An EBS volume can be attached to one, and only one EC2 instance at a time.

As of June 2014, EBS volumes are now available using SSD option in addition to the standard drive option previously offered. There are now three choices for EBS configuration. The options include a Magnetic, a General Purpose SSD, and a Provisioned IOPS SSD option. The cost and performance depends on the type you choose. Prices range from five cents per hour, up to twelve and a half cents per hour for PIOPS SSD, plus some additional fees if choosing Magnetic, or Provisioned IOPS SSD. Unlike Instance storage, you pay for every gigabyte you provision. Even if you have free space on the volume, it's charged as provisioned storage.

EBS volumes are redundant within an AWS Availability Zone (AZ), however, it is strongly recommended you back up volumes in the event of an AZ outage. With EBS, AWS provides a built-in data snapshot functionality for backup. Snapshots are stored in the Amazon Simple Storage Service (S3). Amazon S3 stores multiple copies of data across an entire region and thus achieves 99.999999999 of durability for your snapshots, and other data stored in S3. A snapshot is an image of the entire EBS volume the first time you create a snapshot. Every additional snapshot is an incremental copy of only changed data on the EBS volume. Although the process is very efficient, and can be run while the drive is in use, for heavily utilized EBS volumes you should consider taking snapshots off-peak.

You can also choose EC2 instances that have the EBS-optimized option available. EBS-optimized instances provide a dedicated 500Mb, 1000Mb, and 2000 MB storage NIC for your EBS volumes. There is a small additional fee for using the EBS Optimized feature. Utilizing EBS optimized instances will increase EBS performance, and should be used especially in cases where you choose EBS Provisioned IOPS.

When attaching multiple volumes to your EC2 instance, you can also stripe EBS volumes at the OS level to further increase performance. EBS volumes can also be used to create multi-petabyte Network File Systems.

Security in Brief:

Just like instance storage, you can choose to encrypt your EBS volumes, and launch instances into a VPC. You can use any encryption tools for file systems you wish. Recently, AWS added the option of encryption through their management interface. In this model AWS manage the keys for encryption. You should also consider securing access to your instances via SSL/TLS and configuring your security group(s) with a "least privilege" access approach.

Ideal for:

- EBS volumes are ideal for use cases where the data contained on the volume must persist. As mentioned above, EBS data can persist regardless of the health or lifespan of the EC2 instance it's attached to. In addition to serving as the boot volume for your instance, and storing data that must persist past the life of the EC2 instance, EBS is ideal for running your own relational database. For databases, you should consider using General Purpose SSD, or for even better performance, select the Provisioned IOPS SSD drive option. If it is an option with the EC2 instance option you choose, also consider enabling the "EBS optimized" feature.
- For more information see: http://aws.amazon.com/ebs

Object Storage—Amazon S3 and Amazon Glacier:

The Simple Storage Service (S3) provides massive scale storage for the Internet. You can store any amount of data, and do so in huge object sizes of up to 5 terabytes. Amazon S3 is an object storage solution designed with "Write Once Read Many" architecture. S3 objects are stored in containers called buckets. An AWS account can have up to one hundred buckets. An object is defined as a file, and optionally, any metadata about that file. When you copy data into S3 it stores multiple copies of your files across multiple facilities across a region. With the exception of US Standard, all data will remain in the region you select unless you explicitly move it to another location. All data in S3 is available via a URL, and can be made accessible via various access policies.

S3 includes many built-in tools to help manage large data volumes. For example, you can choose to enable versioning of your data to keep all versions of your files. Logging is simple with S3 and you can easily enable logging when creating an S3 bucket. With logging enabled you can capture what has been accessed when, and by whom. You can choose to set up folders to give the appearance of hierarchy to your users. S3 provides lifecycle management policies to manage the data your store with the service. Using lifecycle policies you can archive and/or delete data based on a date. You can also choose to expire data through the lifecycle policy, which effectively deletes data based on conditions you set.

Amazon Glacier is a highly durable, very inexpensive option for "cold" storage. You can easily move data from S3 to Glacier using the lifecycle policies. You can also move data directly into Glacier via the SDKs and third-party solutions. Unlike S3, the containers in Glacier are called Vaults, and the objects are called Archives. Amazon Glacier is designed to store long-term data that you don't need real-time access to. Data stored in Glacier is charged at a cheaper rate, but to retrieve the data there is a three- to five-hour delay, and a fee (if you retrieve more than 5 percent of your total data stored).

Security in Brief:

You can use client or server side encryption when you store data in S3. The encryption standard is AES 256, and it is easily set up. The encryption of the data is retained if and when you archive data to Glacier. Another way to secure your S3 data is to utilize authorization polices via IAM, Bucket, and ACL policies individually. The policies can be applied independently, or you could use a mix two, or even all three types of policies. You can secure S3 via SSL/TLS. You can also use secure signed URLs with S3.

Ideal for:

- S3 is ideal for static data, website files, and video files. You can even host entire websites in S3. Glacier costs significantly less than S3, and is best for long-term storage of files that you don't need in real time. For example, long-term financial record storage.
- For more information see: S3: http://aws.amazon.com/s3
- Glacier: http://aws.amazon.com/glacier

Content Delivery—Amazon Cloudfront

Cloudfront is Amazon's content distribution network. Cloudfront enables caching across the globe to speed up users' web experience. Cloudfront provides content delivery via fifty-two edge locations around the world (and the number is always growing). The data is stored in servers called origins, and you can have multiple origin servers with a single Cloudfront distribution. Once deployed, your distributions even get a URL that you can point users at for content. You can enable a web distribution, or an RTMP distribution for video content distribution.

Cloudfront can distribute static, dynamic, and streaming data. Customers can utilize multiple origin servers to store content, and multiple Cloudfront distributions to support their applications. You can use Amazon S3, Amazon EC2, or your own servers as origins. When you choose to use S3 as a Cloudfront origin, you get unlimited storage for your files and highly available, secure, and durable storage.

Security in Brief:

You can secure your Cloudfront distributions with ssl/tls and you can encrypt the data in S3 or EC2 to increase your security posture. When you use S3 for your origin(s), you can prevent users from going to directly to the S3 urn using an Origin Access Identity in Cloudfront.

Ideal for:

- Cloudfront is ideal for content delivery of static, dynamic, and streaming content across the globe. You can stream web and video content to some, or all of the Cloudfront edge locations.
- For more information see: http://aws.amazon.com/cloudfront/

Relational Databases—The Amazon Relational Database Service (RDS)

The Relational Database Service is a managed database solution that allows you to focus on your table designs, and your data. In RDS, Amazon manages the patching of your databases and data backup. RDS support Microsoft SQL server, Oracle, Postgres SQL, and MySQL platforms. RDS databases can be backed up via the built-in snapshot functionality provided by AWS.

You can choose single or multi-availability zone (multi-AZ) RDS instance configurations. With a multi-AZ configuration, you ensure that your database is available in the unlikely event of an availability zone outage. AWS handles the fail-over process to the healthy RDS instance in the other availability zones. RDS supports databases ranging in size from five gigabytes up to three terabytes (the new service, Amazon RDS for Aurora is an exception to this size limitation). You can use your existing database tools to work with your data in the databases stored in RDS instances.

Security in Brief:

Amazon RDS provides customer-controlled Security Groups to lock down access to your RDS instances. You can also encrypt the data within your databases in RDS. RDS instances can be secured via SSL and TLS.

Ideal for:

- RDS is ideal for relational databases. They enable you to offload the DBA tasks of backup, database patching and automatic fail-over, and host replacement.
- For more information see: http://aws.amazon.com/rds

NoSQL Databases—Amazon DynamoDB

Amazon Dynamo DB is a hosted, managed NoSQL web service. DynamoDB provides high-speed access to your data in one or more NoSQL tables. You can provision performance based on your application's read and write requirements. DynamoDB stores data in tables as Items that can be up to 400 KB in size. You can store unlimited numbers of items in a DynamoDB table.

DynamoDB is highly available, fault tolerant, and scalable. The DynamoDB web service is designed to withstand the concurrent loss of up to three facilities within an AWS region. It provides high-speed data access through the use of SSD drives. The sharding, scaling and data node fault tolerance is managed by AWS. Much like RDS, DyanmoDB's managed service approach allows you focus on managing the data in your tables while AWS handle the underlying infrastructure.

Security in Brief:

Integrating DynamoDB and Amazon Identity and Access Management allows you to use fine-grained access control to our DynamoDB. You can encrypt your data before storing it in DynamoDB tables, and DynamoDB endpoints can be secured with SSL or TLS.

Ideal for:

- New, cloud-ready applications and existing on-premise NOSQL databases you wish to migrate to AWS.
- For more information see: http://aws.amazon.com/dynamodb

Big Data

Amazon Elastic MapReduce

Elastic MapReduce (EMR) is a web service that hosts Hadoop workloads in AWS. EMR can scale to process and analyze multi-petabytes of data. EMR utilizes EC2 instances for processing, and S3 for storage. EMR handles much of the administration tasks for your Hadoop clusters, so you can focus on the data and the analysis thereof. EMR supports HDFS, hive, and pig as well as other industry standard access methods.

You can provision a Hadoop cluster, and have massive numbers of EC2 nodes to process big data. The EC2 instances spin up to complete the work, and terminate when they are no longer needed, thus saving customers costs. You can obviously run your own Hadoop clusters on EC2, but you should definitely evaluate EMR before doing so.

Security in Brief:

You can encrypt data in your Hadoop cluster. You can launch the EMR cluster in a VPC for tighter security, and you can encrypt access via SSL or TLS.

Ideal for:

- Click log analysis, scientific analysis, or financial modeling applications.
- For more information see: http://aws.amazon.com/emr

Amazon Redshift

Amazon Redshift is a fast, highly scalable, data warehousing solution in the AWS cloud. Redshift provides petabyte scale storage and analysis for big data use cases. It provides data warehousing at a fraction of the cost of traditional data center-based data warehousing solutions. Redshift was designed to use optimized hardware and utilizes a massively parallel architecture. Redshift is a customized implementation of Postgres SQL and operates in columnar architecture, as opposed to the traditional row-based database design.

Multiple nodes are presented to us as Redshift Instances. The management of the nodes is done by AWS, and we can scale up Redshift instances while keeping the data warehouse up and running. Queries are possible using ODBC and JDBC. Redshift integrates well with DynamoDB and S3 for data ingestion and output. We can use standard query tools to work with the data in Redshift.

Security in Brief:

Redshift has security groups to allow for configuration of tight security access. You can encrypt data in Redshift clusters, and launch Redshift into the Amazon VPC, and secure API access via SSL/TLS

Ideal for:

- Business Intelligence, Social Media analysis, scientific analysis, financial modeling.
- For more information see: http://aws.amazon.com/redshift

In-Memory Caching—Amazon Elasticache

Amazon Elasticache is a powerful in-memory caching solution that enables high-speed access to your data. Elasticache supports both Redis and MemcacheD caching systems. AWS manages the nodes in your cluster and handles the automatic node replacement if and when you encounter failed node. You can connect to your nodes via a configuration endpoint.

You can pair Elasticache with Simple Notification Service to receive messages on the Elasticache actions. You could also integrate Elasticache with RDS Read Replicas to serve data to customers quickly while offloading the reads to your primary RDS database. You can shard across cache clusters. The cluster runs in an Availability zone and speeds up data access through caching.

Security in Brief:

Elasticache Security Groups allow you to lock down access to your cache to just those ports you need open for your applications. The clusters can be launched in a VPC in private subnets to further harden security posture.

Ideal for:

- Elasticache is ideal for caching HTML pages, page fragments, and database records to name a few. Elasticache can dramatically increase data access for users.
- For more information see: http://aws.amazon.com/elasticache

Conclusion

In this paper, we've explored the native AWS storage solutions. Remember, our goal is to deliver applications in the cloud in the most efficient, cost-effective, and secure manner. In terms of storage, it's important to understand the characteristics of each AWS storage option. Then you can implement one or more AWS storage services to meet your needs. Often, you'll find that utilizing multiple storage options together will give you the best outcomes. As AWS often notes when referencing storage "one size does not fit all."

Learn More

Learn more about how you can improve productivity, enhance efficiency, and sharpen your competitive edge through AWS training.

AWS Essentials

Architecting on AWS

Architecting on AWS - Advanced Concepts

Systems Operations on AWS

Developing on AWS

Big Data on AWS

Visit **www.globalknowledge.com** or call **1-800-COURSES (1-800-268-7737)** to speak with a Global Knowledge training advisor.

About the Author

Chris Littlefield has been training and consulting in the IT space for the past 18 years. His career has spanned many IT disciplines. Starting as a network engineer, and Microsoft Certified Trainer, Chris consulted and taught throughout the US, and Australia. He’s completed large scale financial IT projects in the US, Australia, India and China. His background includes projects in infrastructure, business intelligence and cloud architecture. Chris began working with the AWS platform as an architect and developer in 2010. He’s been training Amazon Web Services courses for Global Knowledge for the past year.