Technology Overview Study Notes

Andrew Brown

EC2 Cloud Computing

Elastic Compute Cloud (EC2) allows you to launch Virtual Machines (VM). When we launch a Virtual Machine we call it an ”instance”.

What is a Virtual Machine?

A **Virtual Machine** (VM) is an emulation of a physical computer using software.

Server Virtualization allows you to easily **create**, **copy**, **resize** or **migrate** your server.

Multiple **VMs** can run on the same physical server so you can share the cost with other customers.

Imagine if your server or computer was an executable file on your computer

An **Amazon Machine Image (AMI)** is a predefined configuration for a Virtual Machine.

EC2 is considered the backbone of AWS because the majority of AWS services are using EC2 as their underlying servers. eg. S3, RDS, DynamoDB, Lambdas

EC2 is a highly configurable server where you can choose `AMI` that affects options such as:

- The amount of CPUs
- The amount of Memory (RAM)
- The amount of Network Bandwidth
- The Operation System (OS) eg. Windows 10, Ubuntu, Amazon Linux 2
- Attach multiple virtual hard-drives for storage eg. Elastic Block Store (EBS)

Containers

- virtualizing an Operation System (OS) to run multiple workloads on a single OS instance. Containers are generally used in micro-service architecture (when you divide your application into smaller applications that talk to each other)

Elastic Container Service (ECS) - is a container orchestration service that support Docker containers. Launches a cluster of server(s) on EC2 instances with Docker installed. When you need Docker as a Service, or you need to run containers.

Elastic Container Registry (ECR) - is repository for container images. In order to launch a containers you need an image. An image just means a saved copy. A repository just means a storage that has version control.

ECS Fargate - is a serverless orchestration container service. It is the same as ECS expect you pay-on-demand per running container (With ECS you have to keep a EC2 server running even if you have no containers running) AWS manages the underlying server, so you don't have to scale or upgrade the EC2 server.

Elastic Kubernetes Service (EKS) - is a fully managed Kubernetes service. Kubernetes (K8) is an open-source orchestration software that was created by Google and is generally the standard for managing microservices. When you

need to run Kubernetes as a Service.

Serverless

- when the underlying servers are managed by AWS. You don't worry about or configure servers.

AWS Lambdais a serverless functions service. You can run code without provisioning or managing servers. You upload small pieces of code, choose much memory and how long function is allowed to run before timing out. You a charged based on the runtime of the serverless function rounded to the nearest 100ms.

Virtual Machines

- an emulation of a physical computer using software

Amazon LightSale - is the managed virtual server service. It is the "friendly" version of EC2 Virtual Machines

What is Edge Computing?

When you push your computing workloads outside of your networks to run close to the destination location.

eg. Pushing computing to run on phones, IoT Devices, or external servers not within your cloud network.

##What is Hybrid Computing?

When you're able to run workloads on both your on-premise datacenter and AWS Virtual Private Cloud (VPC)

AWS Outposts - is physical rack of servers that you can put in your data center. AWS Outposts allows you to use AWS API and Services such as EC2 right in your

datacenter.

AWS Wavelength - AWS Wavelength allows you to build and launch your applications in a telecom datacenter. By doing this your applications with have ultra-low latency since they will be pushed over a the 5G network and be closest as possible to the end user.

VMWare Cloud on AWS - allows you to manage on-premise virtual machines using VMWare as EC2 instances. The data-center must being using VMWare for Virtualization.

AWS Local Zones - are edge datacenters located outside of an AWS region so you can use AWS closer to end destination. When you need faster computing, storage and databases in populated areas that are outside of an AWS Region

Cost Management

How do we save money?

Capacity Management

How do we meet the demand of traffic and usages though adding or upgrading servers?

EC2 Spot Instances, Reserved Instanced and Savings Plan - Ways to save on computing, by paying up in full or partially, by committing to a yearly contracts or by being flexible about availability and interruption to computing service.

AWS Batch - plans, schedules, and executes your batch computing workloads across the full range of AWS compute services, can utilize Spot Instance to save money.

AWS Compute Optimizer - Suggests how to reduce costs and improve performance by using machine learning to analyze you previous usage history

EC2 Autoscaling Groups (ASGs) - Automatically adds or remove EC2 servers to meet the current demand of traffic. Will save you money and meet capacity since you only run the amount of servers you need.

Elastic Load Balancer (ELB) - Distributes traffic to multiple instance, can re-route traffic from unhealthy instance to healthy instances. Can route traffic to EC2 instances running in different Availability Zones

Elastic Beanstalk (EB) - is for easily deploying web-applications without developers having to worry about setting up and understanding the underlying AWS Services. Similar to Heroku.

The Nitro System

A combination of dedicated hardware and lightweight hypervisor enabling faster innovation and enhanced security. All new EC2 instance types use the Nitro System.

- **Nitro Cards** — specialized cards for VPC, EBS and Instance Storage and controller card
- **Nitro Security Chips** — Integrated into motherboard. Protects hardware resources.
- **Nitro Hypervisor** — lightweight hypervisor Memory and CPU allocation Bare Metal-like performance

Bare Metal Instance

You can launch EC2 instance that have no hypervisor so you can run workloads directly

on the hardware for maximum performance and control. The M5 and R5 EC2 instances run are bare metal.

- **Bottlerocket** is a Linux-based open-source operation system that is purpose-

built by AWS for running containers on Virtual Machines or bare metal hosts

What is High Performance Computing (HPC)?

A cluster of hundreds of thousands of servers with fast connections between each of them with the purpose of boosting computing capacity. When you need a supercomputer to perform computational problems to large to fix on a standard computers or would take to long.

- **AWS ParallelCluster** is an AWS-supported open source cluster management tool that makes it easy for you to deploy and manage High Performance Computing (HPC) clusters on AWS.

Storage Services:

Provide scalable cloud based storage solutions for your workloads on AWS.

- **S3 (Simple Storage Service)** - is an `object storage` service that offers industry-leading scalability, data availability, security, and performance. Think of it as a `"hard drive in the cloud"` with a lot of available space.

- **S3 Glacier** - low cost storage for `archiving and long-term backup` Trade-off: You may have to wait several hours to access data stored here. Use case: for data that you must hold on to but are unlikely to look at often. Example: an enterprise company that must store records for many years under litigation hold.

- **EBS** - `Elastic Block Storage-` is a persitant block storage service. It is a virtual hard drive in the cloud you attach to EC2 instances. You can choose different kinds of hard drives: **SSD, IOPS SSD, Throughput HHD, Cold HHD**

- **EFS** - `Elastic File Storage-` file storage mountable to multiple EC2 instances at the same time

- **Storage Gateway** - hybrid cloud storage with local caching. Expand your on-premises storage capacity into the cloud.

- **Snowball** - physically migrate large quantities of data to AWS via a mobile computer suitcase - `50-80 TB`

- **Snowball edge** - A version of Snowball for even larger sets of data - `100 TB`

- **Snowmobile** - Shipping container, pulled by a semi-trailer truck for the largest of migrations - `100 PB`

NoSQL Database Services

DynamoDB - is a serverless NoSQL key/value and document database. It is designed to scale to billions of records with guaranteed consistent data return in at least a second. You don't have to worry about managing shards!

- DynamoDB is AWS's flagship database service meaning whenever we think of a database service that just scales, is cost effective and very fast we should think DynamoDB

DocumentDB - is a NoSQL document database that is "MongoDB compatible"

- MongoDB is very popular NoSQL among developers. There were open-source licensing issues around using open-source MongoDB, so AWS got around it by just building their own MongoDB database.

Amazon Keyspaces - is a fully managed Apache Cassandra database. Cassandra is an open-source NoSQL key/value database similar to DynamoDB in that is columnar store database but has some additional functionality. When you want to use Apache Casandra.

Relational Database Services

Relational Database Service (RDS) - is a relational database service that supports multiple SQL engines. Relational is synonymous with SQL and Online Transactional Processing (OLTP). Relational database are the most commonly used type of database among tech companies and start-ups.

RDS Supports the following SQL Engines:

- **MySQL** – The most popular open-source SQL database that was purchased and now owned by Oracle.
- **MariaDB** – When Oracle bought MySQL a fork (copy) of MySQL was made under a different open-source license.
- **Postgres (PSQL)** – Most popular open-source SQL database among developers. Has rich-features over MySQL but at added complexity
- **Oracle** – Oracle's proprietary SQL database. Well used by Enterprise companies. You have to buy a license to use it.
- Microsoft SQL Server – Microsoft's proprietary SQL database. You have to buy a license to use it.
- Aurora – Fully managed database.

Aurora - is a fully managed database of either MySQL (5x faster) and PSQL (3x faster) database. When you want a highly available, durable, scalable and secure relational database for Postgres or MySQL

Aurora Serverless - is the serverless on-demand version of Aurora. When you want "most" of the benefits of Aurora but can trade to have cold-starts or you don't have lots of traffic demand

RDS on VMware - allows you to deploy RDS supported engines to on an-premise data-center. The datacenter must be using VMware for server virtualization. When you want databases managed by RDS on your own datacenter

Other Database Services

Redshift - is a petabyte-size data-warehouse. Data-warehouses are for Online Analytical Processing (OLAP) Data-warehouses can be expensive because they are keeping data "hot". Meaning that we can run a very complex query and a large amount of data and get that data back very fast. When you want to quickly generate analytics or reports from a large amount of data.

ElastiCache - is a managed database of the in-memory and caching open-source databases Redis or Memcached. When you need to improve the performance of application by adding a caching layer in-front of web-server or database.

Neptune - is a managed graph database. Data is represented as interconnected nodes. When you need to understand the connections between data eg. Mapping Fraud Rings or Social Media relationships.

Amazon Timestreams - is a fully managed time series database. Think of devices that send lots of data that are time-sensitive such as IoT devices. When you need to measure how things change over time.

Amazon Quantum Ledger Database - is a fully managed ledger database that provides transparent, immutable and cryptographically variable transaction logs.

Database Migration Service (DMS - is database migration service. You can us it to migrate from:

- on-premise database to AWS
- from two database in different or the same AWS accounts using different

SQL engines

- from an SQL to NoSQL database

Windows on AWS

Windows Servers on EC2 - You can select from a number of Windows Server versions including the latest version, Windows Server 2019

SQL Server on RDS - You can select from a number of SQL Server database versions

AWS Directory Service - lets you run Microsoft Active Directory (AD) as a managed service

AWS License Manager makes it easier to manage your software licenses from software vendors such as Microsoft.

Amazon FSx for Windows File Server - is a fully managed scalable storage built for Windows.

AWS Software Development Kit (SDK) - allows you to write code in your favorite language to interact with AWS API. The SDK supports .NET a language favorite for Windows Developers

Amazon WorkSpaces allows you to run a virtual desktop. You can launch a Windows 10 desktop to a provide secure and durable workstation that is accessible from wherever you have an internet connection.

AWS Migration Acceleration Program (MAP) - for Windows is a migration methodology from moving large enterprise. AWS has Amazon Partners that specialize in providing professional services for MAP.

Cloud Native Networking Services -- Key definitions:

Region - the geographical location of your network

Availability Zone (AZ) - a data center containing your AWS resources

Virtual Private Cloud (VPC)- a logically isolated section of the AWS Cloud where you can launch AWS resources

Internet Gateway- enables access to the Internet for your VPC

Route Tables- determines where network traffic from your subnets are directed

NACLs- *Network Access Control Lists.* Act as a firewall at the subnet level

Security Groups- Act as firewall at the instance level

Subnets- a logical partition of an IP network into multiple, smaller network segments

Availability Zones:

Availability Zones are the **data centers** where you launch your AWS resources into

Each AZs is associated with a specific `region`

Key VPC Components:

A virtual private cloud (VPC) network is your own personal isolated section of the AWS cloud.

A `route table` contains a set of rules (called routes), that are used to determine where network traffic from your subnet or gateway is directed.

Internet Gateway- Allows you to grant internet access to resources inside of

your VPC. But you also need a **route table** which routes the traffic from the VPC network out to the IGW

You can think of it as a door from your VPC outward.

Enterprise Hybrid Networking Services

Direct Connect - is a dedicated Gigabit network connection from your premises location to AWS. Provides a direct fiber optic cable running straight to the AWS network

VPN - establishes a secure connection to your AWS network

- Site-to-Site VPN - connecting your on-premise to your AWS network
- Client VPN - connecting a Client (ie users laptop) to your AWS network

PrivateLinks (VPC Interface Endpoints) - keeps traffic within the AWS network and not traverse the internet to keep traffic is secure.

Virtual Private Cloud & Subnets

Virtual Private Cloud (VPC)- a logically isolated section of the AWS Cloud where you can launch AWS resources

Subnets- a logical partition of an IP network into multiple, smaller network segments

- Subnets need to have a smaller CIDR range than to the VPC represent their portion. eg Subnet CIDR Range 10.0.0.0/24 = 256 IP Addresses

Public vs Private Subnets:

Public subnets are generally used for placing resources which are accessible on the internet

`Private subnets` are used when you need resources to be more secured and only accessible through tightly filtered traffic into the subnet

Accounts and Organizations

- **Organizations** - allow you to centrally manage billing, control access, compliance, security, and share resources across your AWS accounts.
- **Root Account User** - a single sign-in identity that has complete access to all AWS services and resources in an account. Each account has a Root Account User.
- **Organization Units** - are a group of AWS accounts within an organization which can also contain other organizational units - creating a hierarchy.
- **Service Control Polices** - give central control over the allowed permissions for all accounts in your organization, helping to ensure your accounts stay within your organization's guidelines.

Key Points

- In AWS you can have more than one account managed through a single account using Organizations
- Organizations allow you to setup consolidated billing where 1 account pays the AWS bill for all
- The payer account is the root level account in an Organization
- You can create isolate AWS accounts for different teams under the payer account - and place them inside Organizational Units (OU)
- The separation of accounts into OUs allows you to set customized permission boundaries on the accounts using Service Control Policies (SCPs)

Provisioning Services

`Provisioning` is the allocation or creation of resources and services to a customer. AWS Provisioning Services are responsible for setting up and then managing those AWS Services

Types of Services

- **Elastic Beanstalk** - an easy-to-use service for deploying and scaling web applications and services developed with Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker on familiar servers such as Apache, Nginx, Passenger, and Internet Information Services (IIS).

- **OpWorks** - configuration management service that provides managed instances of **Chef** and **Puppet**

- **CloudFormation** - lets you deploy your cloud resources using infrastructure-as-code with **JSON** and **YAML** templates

- **AWS QuickStarts** - pre-made packages that can launch and configure your AWS compute, network, storage, and other services required to deploy a workload on AWS

- **AWS Marketplace** - a digital catalogue containing **thousands** of software listings from independent software vendors you can use to find, buy, test, and deploy software.

- **AWS Amplify** - is a mobile and web-application framework, that will provision multiple AWS services as your backend.

Logging Services

Amazon Web Services (AWS) provides service-specific operational metrics and log files to give customers insight into how the service is operating.

CloudTrail

CloudTrail - logs all `API calls`(SDK, CLI) between various `AWS services`

Questions that CloudTrail can answer:

Who create this bucket? - detect developer mis-configuration

Who spun up that expensive EC2 instance? - Detect malicious actors

Who launched this SageMaker notebook? - Automate responses

###CloudWatch **CloudWatch** - is a collection of multiple services

- CloudWatch **Logs** : Performance data about AWS Services eg. CPU Utilization, Memory, Network in Application Logs eg. Rails, Nginx Lambda Logs
- CloudWatch **Metrics**: Represents a time-ordered set of data points. A variable to monitor
- CloudWatch **Events**: trigger an event based on a condition eg. every hour take a snapshot of the server
- CloudWatch **Alarms**: triggers notifications based on metrics
- CloudWatch **Dashboard**: create visualizations based on metrics

X-Ray

AWS X-Ray is a distributed tracing system. You can use it to pinpoint issues with your microservices. See how data moves from one app to another, how long it took to move, and if it failed to move forward.

Enterprise Integration

Often a mixture

Enterprise environments are often a mix of cloud, on-premises data centers, and edge locations. Hybrid cloud architectures help organizations integrate their on-premises and cloud operations to support a broad spectrum of use cases using a common set of cloud services, tools, and APIs across on-premises and cloud environments.
Hybrid Cloud with AWS

Services

Direct Connect - is a dedicated Gigabit network connection from your premises location to AWS. Provides a direct fiber optic cable running straight to the AWS network

VPN - establishes a `secure` connection to your AWS network

- `Site-to-Site VPN` - connecting your on-premise to your AWS network
- `Client VPN` - connecting a Client (ie users laptop) to your AWS network

Storage Gateway - is a hybrid storage service that enables your on-premises applications to use AWS cloud storage. You can use this for backup and archiving, disaster recovery, cloud data processing, storage tiering, and migration.

Active Directory - AWS Directory Service for Microsoft Active Directory also know as AWS Managed Microsoft AD- enables your directory-aware workloads and AWS resources to use managed Active Directory in the AWS Cloud.

Business Centric Services

Amazon Connect - `Call center` as a self-service - cloud-based contact center service that makes it easy for any business to deliver better customer service at lower cost.

WorkSpaces - `Virtual Remote Desktops` - a fully managed, secure cloud desktop service. You can use Amazon WorkSpaces to provision either Windows or Linux desktops in just a few minutes and quickly scale to provide thousands of desktops to workers across the globe.

WorkDocs - a content creation and collaboration service. Easily create, edit, and share content saved centrally in AWS

Chime - AWS platform for `online meetings, video conferencing`, and business calling which elastically scales to meet your capacity needs.

WorkMail - Managed `business email`, contacts, and calendar service with support for existing desktop and mobile email client applications

Pinpoint - Marketing campaign management system you can `use for sending targeted email`, SMS, push notifications, and voice messages.

SES - Simple Email Service- A cloud based email sending service designed fro marketers and application developers to `send marketing, notifications, and emails.`

QuickSight - A Business Intelligence (BI) service. Connect multiple datasource and quickly visualize data in the form of graphs with little to no programming knowledge.

Application Integration Services

Simple Notification Service (SNS) - a pub-sub messaging system. Sends notifications via various formats such as Plain-text Email, HTTP/s (webhooks)

SMS (text messages), SQS and Lambda. Push messages which then are sent to subscribers

Simple Queue Service (SQS) is a queueing messaging service. Send events to a queue. Other applications pull the queue for messages. Commonly used for background jobs.

Step Functions - is a state machine service. It coordinate multiple AWS services into serverless workflows. Easily share data among Lambdas. Have a group of lambdas wait for each other. Create logical steps. Also works with Fargate Tasks.

EventBridge (CloudWatch Events) - is a serverless event bus that makes it easy to connect applications together from your own application, third-party services and AWS services.

Kinesis - is a real-time streaming data service. Create Producers which send data to a stream. Multiple Consumers can consume data within a stream. Use for real-time analytics, click streams, ingesting data from a fleet of IOT Devices

Amazon MQ - is a managed message broker service that uses Apache ActiveMQ

Managed Kafka Service (MSK) - a fully managed Apache Kafka service. Kafka is an open-source platform for building real-time streaming data pipelines and applications. Similar to Kinesis but more robust

API Gateway - is a fully-managed service for developers to create, publish, maintain, monitor, and secure APIs. You can create API endpoints and route them to AWS services.

AppSync - is a fully managed GraphQL service. GraphQL is an open-source agnostic query adaptor that allows you to query data from many different data sources.

What is Big Data?

A term used to describe massive volumes of structured/unstructured data that is so large it is difficult to move and process using traditional database and software techniques.

Amazon Athena - is a serverless interactive query service. It can take a bunch of CSV or JSON files in a S3 Bucket and load them into temporary SQL tables so you can run SQL queries. When you want to query CSV or JSON files

Amazon CloudSearch - is a fully managed full-text search service. When you want add search to your website

Amazon Elasticsearch Service (ES) - is a managed Elasticsearch cluster. Elasticsearch is a open-source full-text search engine. It is more robust than CloudSearch but requires more server and operational maintaince.

Amazon Elastic MapReduce (EMR) - is for data processing and analysis. Its can be used for creating reports just like Redshift, but is more suited when you need to transform unstructured data into structured data on the fly.

Kinesis Data Streams - is a real-time streaming data service. Create Producers which send data to a stream. Multiple Consumers can consume data within a stream. Use for real-time analytics, click streams, ingesting data from a fleet of IOT Devices

Kinesis Firehose - is serverless and a simpler version of Data Streams, You pay-on-demand based on how much data is consumed through the stream and you don't worry about the underlying servers.

Amazon Kinesis Data Analytics - allows you to run queries against data that is flowing through your real-time stream so you can create reports and analysis on emerging data.

Managed Kafka Service (MSK) - a fully managed Apache Kafka service. Kafka is an open-source platform for building real-time streaming data pipelines and applications. It is similar to Kinesis but with more robust functionalities

Redshift - is a petabyte-size data-warehouse. Data-warehouses are for Online Analytical Processing) OLAP Data-warehouses can be expensive because they are keeping data “hot”. Meaning that we can run a very complex query and a large amount of data and get that data back very fast. When you to quickly generate analytics or reports from a large amount of data.

What is Serverless?

When the underlying servers, infrastructure and Operating System (OS) is taken care of by the Cloud Service Provider (CSP).

Serverless is generally be default highly available, scalable and cost-effective. You pay for what you use.

DynamoDB - is a serverless NoSQL key/value and document database. It is designed to scale to billions of records with guaranteed consistent data return in at least a second. You don’t have to worry about managing shards!

Simple Storage Service (S3) - is a serverless object storage service. You can upload very large and an unlimited amount of files. You pay for what you store. You don’t worry about the underlying file-system, or upgrading the disk size.

ECS Fargate - is serverless orchestration container service. It is the same as ECS expect you pay-on-demand per running container (With ECS you have to keep a EC2 server running even if you have no containers running) AWS manages the underlying server, so you don’t have to scale or upgrade the EC2 server.

AWS Lambda - is a serverless functions service. You can run code without provisioning or managing servers. You upload small pieces of code, choose much memory and how long function is allowed to run before timing out. You a charged

based on the runtime of the serverless function rounded to the nearest 100ms.

Step Functions - is a state machine service. It coordinate multiple AWS services into serverless workflows. Easily share data among Lambdas. Have a group of lambdas wait for each other. Create logical steps. Also works with Fargate Tasks.

AI, ML, and DL Services

What is Artificial Intelligence (AI)?

Machines that perform jobs that mimic human behavior

What is Machine Learning (ML)?

Machines that get better at a task without explicit programming

What is Deep Learning (DL)?

Machines that have an artificial neural network inspired by the human brain to solve complex problems.

Amazon SageMaker - is a fully managed service to build, train, and deploy machine learning models at scale

Amazon Augmented AI - human-intervention review service. When SageMaker's uses machine Learning to make a prediction is not confident it has the right answer queue up the predication for human review.

Amazon CodeGuru - is machine-learning code analysis service. CodeGuru performs code-reviews and will suggest changes to improve the quality of code. It can visual code profiles (show the internals of your code) to pinpoint performance.

Amazon Lex - is a conversion interface service. With Lex you can build voice and text chatbots

Amazon Polly - is a text-to-speech service. Upload your text and an audio file spoken by synthesized voice is generated.

Amazon Transcribe - is a speech-to-text service. Upload your audio file and it is converted

Amazon Textract - and OCR (extract text from scanned documents) service. When you have paper forms and you want to digitally extract the data.

Amazon Forecast - is a time-series forecasting service. Forecast business outcomes such as product demand, resource needs or financial performance.

AWS DeepRacer - a toy race car that can be powered with machine-learning to perform autonomous driving.