DIGITAL TALENT SCHOLARSHIP 2020

digitalent.kominfo.go.id

AWS Compute

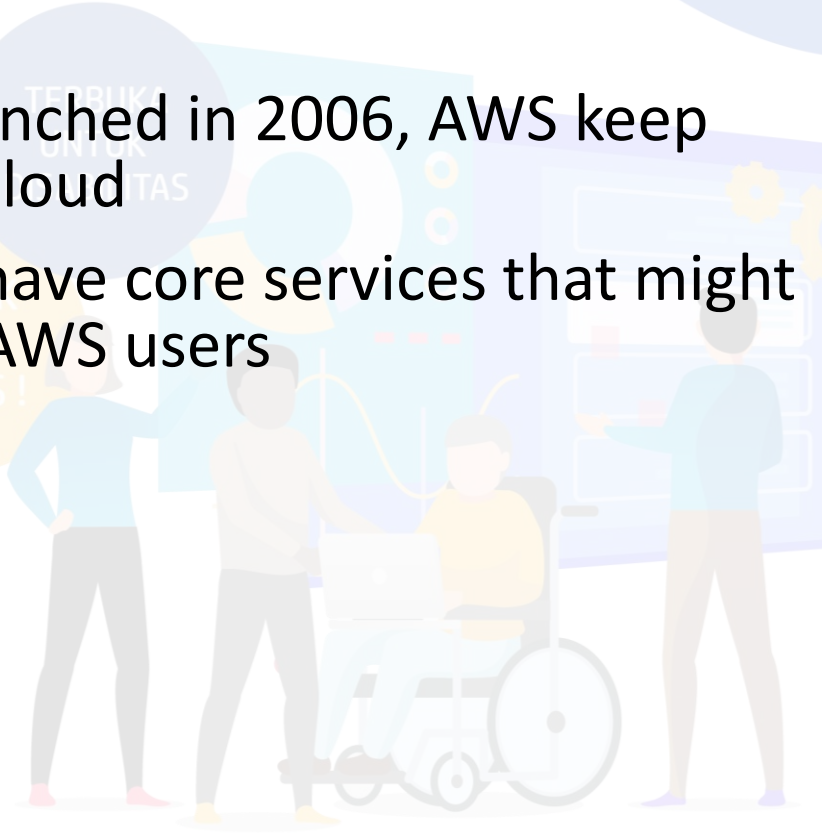
Agenda

- AWS Core Services
- Introduction to AWS Computing Services
- Instance Types in AWS
- Storage Volume in AWS
- Pricing Models in AWS
- Case Study

AWS Core Services

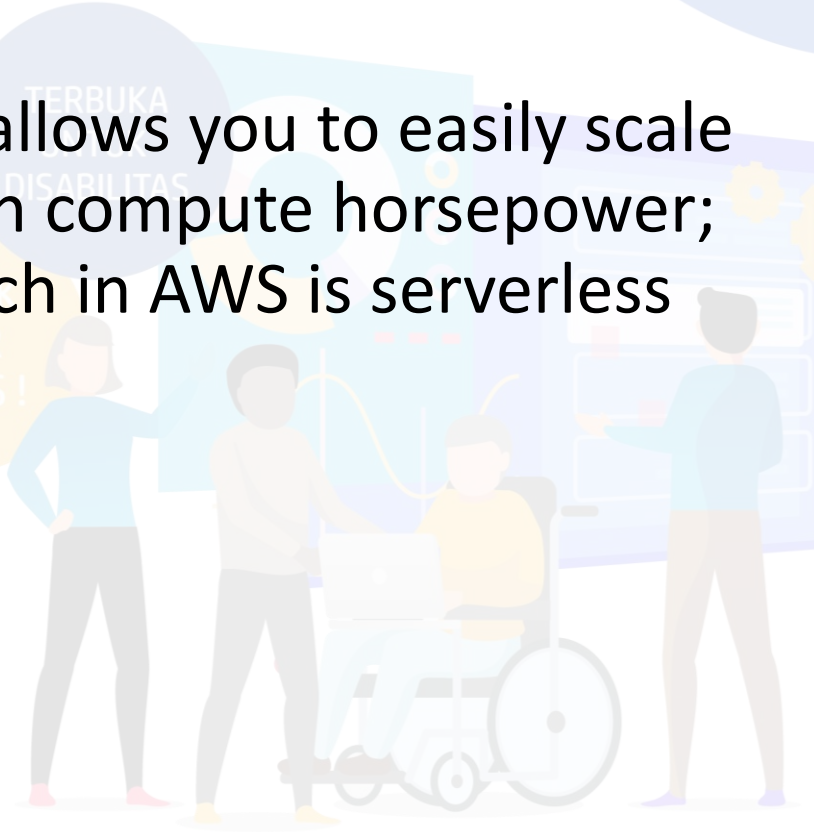
AWS Services

- Since first AWS service (EC2) launched in 2006, AWS keep adding on more services on its cloud
- Among all of AWS services, we have core services that might get frequently used by most of AWS users
- Those services are grouped as :
  - Compute
  - Storage
  - Database
  - Networking
  - Management
  - Security
  - Application Integration



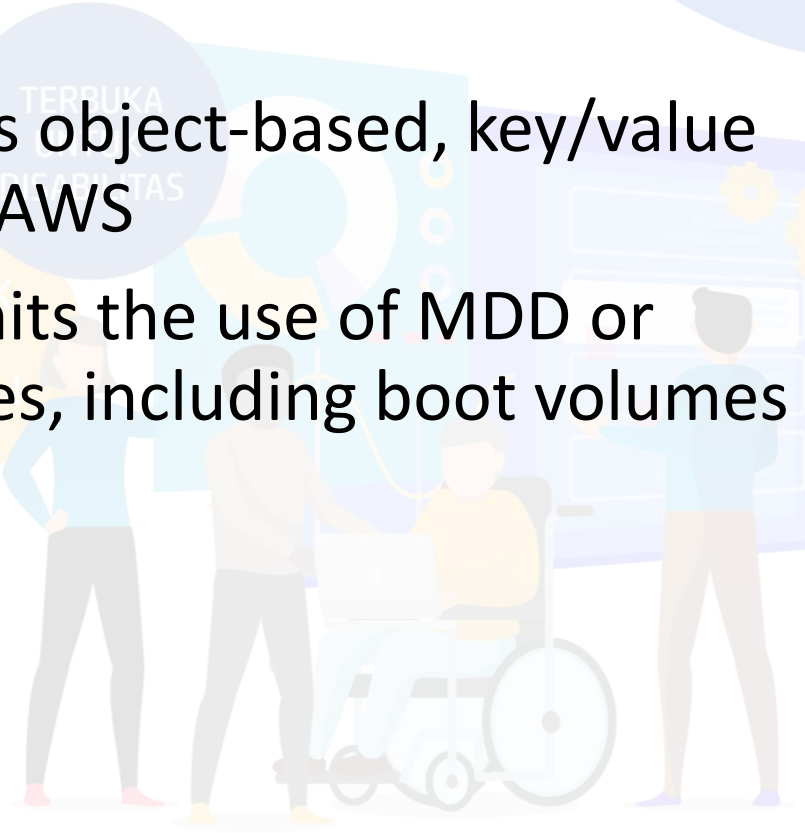
AWS Compute

- **EC2** – Elastic Compute Cloud allows you to easily scale virtual machines for your main compute horsepower; note that an opposite approach in AWS is serverless compute with Lambda



AWS Storage

- **S3** – Simple Storage Services is object-based, key/value storage for many purposes in AWS
- **EBS** – Elastic Block Store permits the use of MDD or SDD storage for many purposes, including boot volumes for EC2 instances

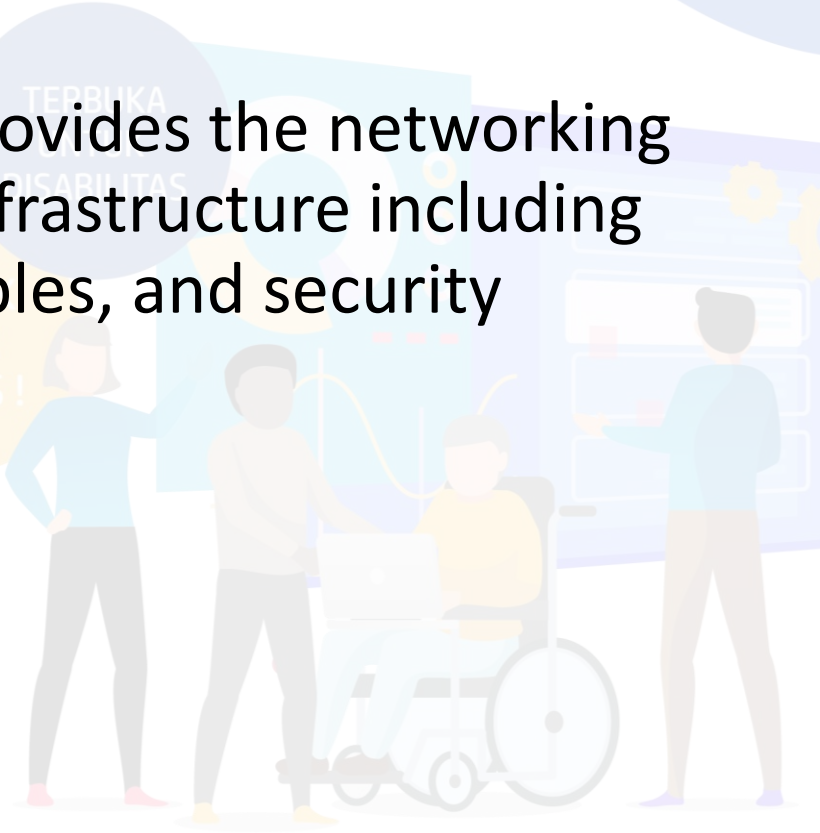


AWS Database

- **RDS** – Relational Database Service allows you to host many database types in the cloud; this includes Oracle, MS SQL Server, and even Amazon's own Aurora
- **DynamoDB** – a NoSQL database option in the cloud that performs blazing performance and on-demand scalability
- **ElastiCache** – creates in-memory caches for impressive performance interaction; this service also supports open standards in caching

AWS Networking

- **VPC** – Virtual Private Cloud provides the networking components needed for an infrastructure including subsets, gateways, routing tables, and security mechanisms



AWS Management

- **CloudWatch** – Permits the monitoring of key services; uses metrics and alarms for a familiar monitoring approach
- **CloudTrail** – Permits the tracking of potentially all the API calls to AWS; this allows you detailed analysis of all events no matter the source – Web Console, CLI, SDK, etc.

AWS Security

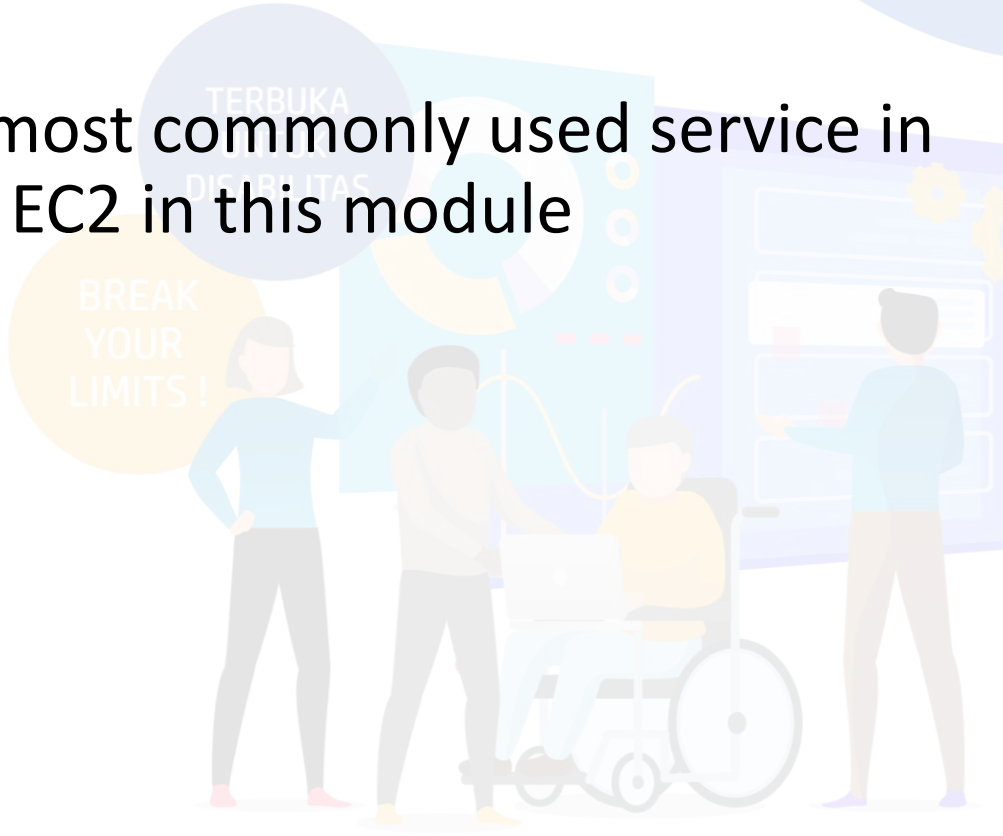
- **IAM** – Identity and Access Management allows the creation of users, groups, and roles for interacting securely with AWS

AWS Application Integration

- **SNS** – the Simple Notification Service allows the generation of email and text notifications based on AWS events
- **SQS** – the Simple Queue Service assists you in decoupling components and queuing messages between these components; this service helps the use of micro services for your processing needs

Most Commonly Used Service

- Since EC2 is one of the most commonly used service in AWS, we will talk about EC2 in this module



Introduction to AWS Compute Services

AWS Compute Services

- AWS has many compute services that we can use
  - EC2
  - EC2 AutoScaling
  - Elastic Container Registry
  - Elastic Container Service
  - Elastic Kubernetes Service
  - Lightsail
  - Batch
  - Elastic Beanstalk
  - Fargate
  - Lambda
  - Serverless Application Repository
  - Outposts
  - VMware Cloud on AWS
- In this module, we will talk about **EC2, ECS, and Lambda**

AWS Compute Services – EC2

- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.
- The Amazon EC2 simple web service interface allows you to obtain and configure capacity with minimal friction.
- Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.

AWS Compute Services – ECS

- Amazon Elastic Container Service (Amazon ECS) is a highly scalable, high-performance container orchestration service that supports Docker containers and allows you to easily run and scale containerized applications on AWS.
- Amazon ECS eliminates the need for you to install and operate your own container orchestration software, manage and scale a cluster of virtual machines, or schedule containers on those virtual machines.
- With simple API calls, you can launch and stop Docker-enabled applications, query the complete state of your application, and access many familiar features such as IAM roles, security groups, load balancers, Amazon CloudWatch Events, AWS CloudFormation templates, and AWS CloudTrail logs.

AWS Compute Services – Lambda

- AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume—there is no charge when your code is not running.
- With Lambda, you can run code for virtually any type of application or backend service—all with zero administration. Just upload your code, and Lambda takes care of everything required to run and scale your code with high availability.
- You can set up your code to automatically trigger from other AWS services, or you can call it directly from any web or mobile app.

More About EC2

- Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) cloud.
- Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster.
- You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage.
- Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic.
- EC2 have many advantages. A few of the advantages include being highly scalable (one can choose instances with more RAM, CPU etc), they are easy to start and stop (outside the free tier, customers pay for what they use), and they allow for the selection of different platforms (operating systems).

EC2 Features

- Virtual computing environments
- Preconfigured templates for your instances
- Various configurations of CPU, memory, storage, and networking
- Secure login information for your instances using key pairs
- Storage volumes for temporary data
- Persistent storage volumes for your data
- Multiple physical locations for your resources
- A firewall that enables you to specify the protocols, ports, and source IP ranges that can reach your instances
- Static IPv4 addresses for dynamic cloud computing
- Metadata
- Virtual networks you can create that are logically isolated from the rest of the AWS cloud

EC2 Instance Types

Instance Types

- Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.
- Instance types comprise varying combinations of CPU, memory, storage, and networking capacity and give you the flexibility to choose the appropriate mix of resources for your applications.
- Each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.

Instance Types

- Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.
- Instance types comprise varying combinations of CPU, memory, storage, and networking capacity and give you the flexibility to choose the appropriate mix of resources for your applications.
- Each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.
- Categorized into :
  - General Purpose
  - Compute Optimized
  - Memory Optimized
  - Accelerated Computing
  - Storage Optimized

What's Your Use Case?

	General Purpose	Compute Optimized	Memory Optimized	Accelerated Computing	Storage Optimized
Instance Types	T3, T2, M5, M5A, M4	C5, C4	R5, R4, X1e, X1,L, z1d, High Memory Instances	P3, P2, G3, F1	H1, I3, D2
Use Case	Broad	High performance	In-memory databases	Machine learning	Distributed file systems

Instance Types – General Purpose

- General purpose instances provide a balance of compute, memory and networking resources, and can be used for a variety of diverse workloads.
- **Use Cases:**
  - Scale-out workloads such as web servers, containerized microservices, caching fleets, and distributed data stores, as well as development environments

Instance Types – General Purpose Examples : T3

- T3 instances offer a balance of compute, memory, and network resources and are designed for applications with moderate CPU usage that experience temporary spikes in use

Instance	vCPU*	CPU Credits/hour	Mem (GiB)	Storage	Network Performance (Gbps)
t3.nano	2	6	0.5	EBS-Only	Up to 5
t3.micro	2	12	1	EBS-Only	Up to 5
t3.small	2	24	2	EBS-Only	Up to 5
t3.medium	2	24	4	EBS-Only	Up to 5
t3.large	2	36	8	EBS-Only	Up to 5
t3.xlarge	4	96	16	EBS-Only	Up to 5
t3.2xlarge	8	192	32	EBS-Only	Up to 5

Instance Types – Compute Optimized

- Compute Optimized instances are ideal for compute bound applications that benefit from high performance processors.
- **Use Cases:**
  - Batch processing workloads, media transcoding, high performance web servers, high performance computing (HPC), scientific modeling, dedicated gaming servers and ad server engines, machine learning inference and other compute intensive applications.

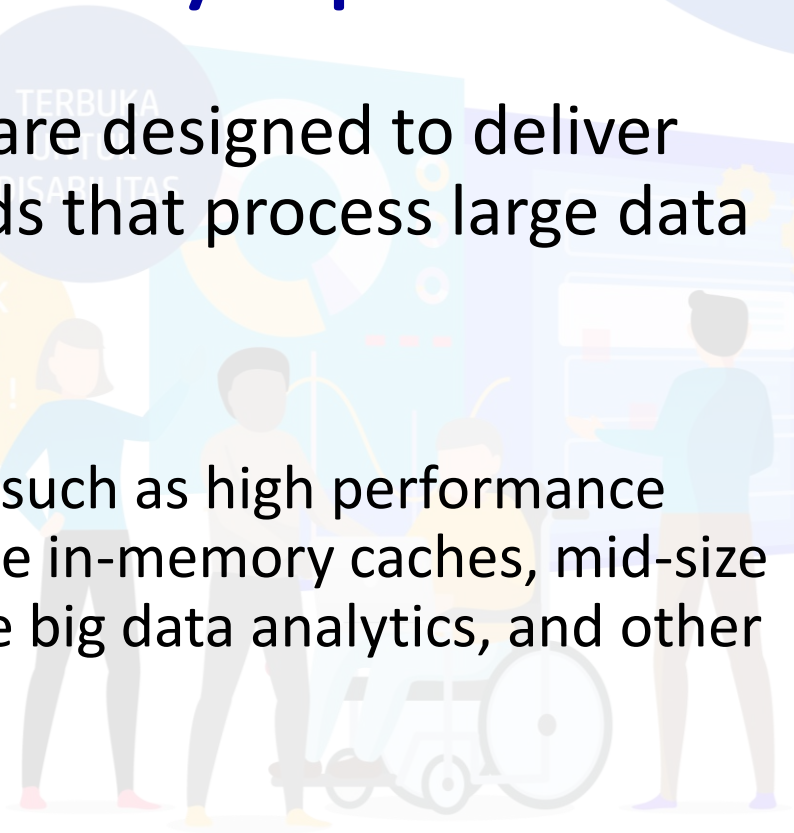
Instance Types – Compute Optimized Examples : C5

- C5 instances are optimized for compute-intensive workloads and deliver cost-effective high performance at a low price per compute ratio.

Model	vCPU	Memory (GiB)	Instance Storage (GiB)	Network Bandwidth (Gbps)	EBS Bandwidth (Mbps)
c5.large	2	4	EBS-Only	Up to 10	Up to 4,750
c5.xlarge	4	8	EBS-Only	Up to 10	Up to 4,750
c5.2xlarge	8	16	EBS-Only	Up to 10	Up to 4,750
c5.4xlarge	16	32	EBS-Only	Up to 10	4,750
c5.9xlarge	36	72	EBS-Only	10	9,500
c5.12xlarge	48	96	EBS-Only	12	9,500
c5.18xlarge	72	144	EBS-Only	25	19,000
c5.24xlarge	96	192	EBS-Only	25	19,000
c5.metal	96	192	EBS-Only	25	19,000
c5d.large	2	4	1 x 50 NVMe SSD	Up to 10	Up to 4,750
c5d.xlarge	4	8	1 x 100 NVMe SSD	Up to 10	Up to 4,750
c5d.2xlarge	8	16	1 x 200 NVMe SSD	Up to 10	Up to 4,750
c5d.4xlarge	16	32	1 x 400 NVMe SSD	Up to 10	4,750
c5d.9xlarge	36	72	1 x 900 NVMe SSD	10	9,500
c5d.12xlarge	48	96	2 x 900 NVMe SSD	12	9,500
c5d.18xlarge	72	144	2 x 900 NVMe SSD	25	19,000
c5d.24xlarge	96	192	4 x 900 NVMe SSD	25	19,000
c5d.metal	96	192	4 x 900 NVMe SSD	25	19,000

Instance Types – Memory Optimized

- Memory optimized instances are designed to deliver fast performance for workloads that process large data sets in memory.
- **Use Cases:**
  - Memory intensive applications such as high performance databases, distributed web scale in-memory caches, mid-size in-memory databases, real time big data analytics, and other enterprise applications.



Instance Types – Memory Optimized Example : R5

- R5 instances deliver 5% additional memory per vCPU than R4 and the largest size provides 768 GiB of memory. In addition, R5 instances deliver a 10% price per GiB improvement and a ~20% increased CPU performance over R4.

Instance	vCPU	Memory (GiB)	Instance Storage (GiB)	Networking Performance (Gbps)	EBS Bandwidth (Mbps)
r5.large	2	16	EBS-Only	up to 10	Up to 4,750
r5.xlarge	4	32	EBS-Only	up to 10	Up to 4,750
r5.2xlarge	8	64	EBS-Only	up to 10	Up to 4,750
r5.4xlarge	16	128	EBS-Only	up to 10	4,750
r5.8xlarge	32	256	EBS-Only	10	6,800
r5.12xlarge	48	384	EBS-Only	10	9,500
r5.16xlarge	64	512	EBS Only	20	13,600
r5.24xlarge	96	768	EBS-Only	25	19,000
r5.metal	96*	768	EBS-Only	25	19,000
r5d.large	2	16	1 x 75 NVMe SSD	up to 10	Up to 4,750
r5d.xlarge	4	32	1 x 150 NVMe SSD	up to 10	Up to 4,750
r5d.2xlarge	8	64	1 x 300 NVMe SSD	up to 10	Up to 4,750
r5d.4xlarge	16	128	2 x 300 NVMe SSD	up to 10	4,750
r5d.8xlarge	32	256	2 x 600 NVMe SSD	10	6,800

Instance Types – Accelerated Computing

- Accelerated computing instances use hardware accelerators, or co-processors, to perform functions, such as floating point number calculations, graphics processing, or data pattern matching, more efficiently than is possible in software running on CPUs.
- **Use Case :**
  - Machine learning inference for applications like adding metadata to an image, object detection, recommender systems, automated speech recognition, and language translation. It also provide a very cost-effective platform for building and running graphics-intensive applications, such as remote graphics workstations, video transcoding, photo-realistic design, and game streaming in the cloud.

Instance Types – Accelerated Computing Examples : G4

- G4 instances are designed to help accelerate machine learning inference and graphics-intensive workloads.

	Instance	GPUs	vCPU	Mem (GiB)	GPU Memory (GiB)	Instance Storage (GB)	Network Performance (Gbps)
Single GPU VMs	g4dn.xlarge	1	4	16	16	125	Up to 25
	g4dn.2xlarge	1	8	32	16	225	Up to 25
	g4dn.4xlarge	1	16	64	16	225	Up to 25
	g4dn.8xlarge	1	32	128	16	1x900	50
	g4dn.16xlarge	1	64	256	16	1x900	50
Multi GPU VMs	g4dn.12xlarge	4	48	192	64	1x900	50
	g4dn.metal	8	96	384	128	2x900	100

Instance Types – Storage Optimized

- Storage optimized instances are designed for workloads that require high, sequential read and write access to very large data sets on local storage. They are optimized to deliver tens of thousands of low-latency, random I/O operations per second (IOPS) to applications.
- **Use Case :**
  - NoSQL databases (e.g. Cassandra, MongoDB, Redis), in-memory databases (e.g. Aerospike), scale-out transactional databases, data warehousing, Elasticsearch, analytics workloads.

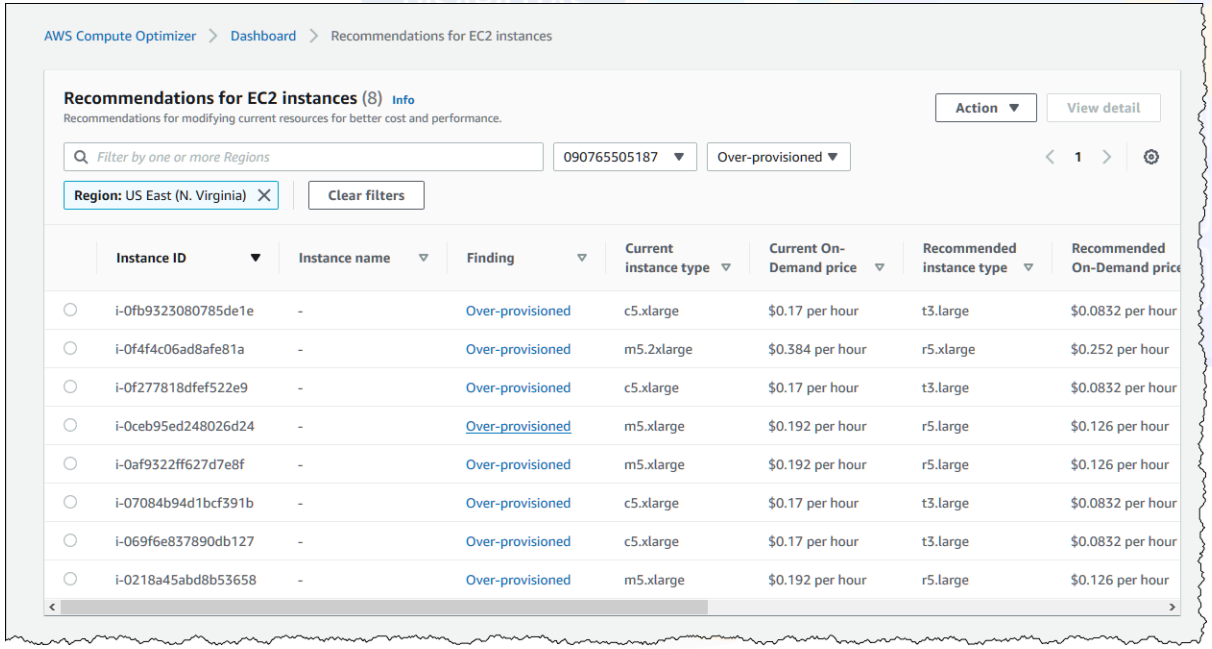
Instance Types – Storage Optimized Examples : I3en

- This instance family provides dense Non-Volatile Memory Express (NVMe) SSD instance storage optimized for low latency, high random I/O performance, high sequential disk throughput, and offers the lowest price per GB of SSD instance storage on Amazon EC2. I3en also offers Bare Metal instances (i3en.metal), powered by the Nitro System, for non-virtualized workloads, workloads that benefit from access to physical resources, or workloads that may have license restrictions.

Instance	vCPU	Mem (GiB)	Local Storage (GB)	Network Bandwidth
i3en.large	2	16	1 x 1,250 NVMe SSD	Up to 25 Gbps
i3en.xlarge	4	32	1 x 2,500 NVMe SSD	Up to 25 Gbps
i3en.2xlarge	8	64	2 x 2,500 NVMe SSD	Up to 25 Gbps
i3en.3xlarge	12	96	1 x 7,500 NVMe SSD	Up to 25 Gbps
i3en.6xlarge	24	192	2 x 7,500 NVMe SSD	25 Gbps
i3en.12xlarge	48	384	4 x 7,500 NVMe SSD	50 Gbps
i3en.24xlarge	96	768	8 x 7,500 NVMe SSD	100 Gbps
i3en.metal	96	768	8 x 7,500 NVMe SSD	100 Gbps

Recommendation for Instance Types

- AWS provides Amazon EC2 instance recommendations to help you improve performance, save money, or both, by using features powered by **AWS Compute Optimizer**.
- You can use these recommendations to decide whether to move to a new instance type.



AWS Storage Volumes

Storage for EC2

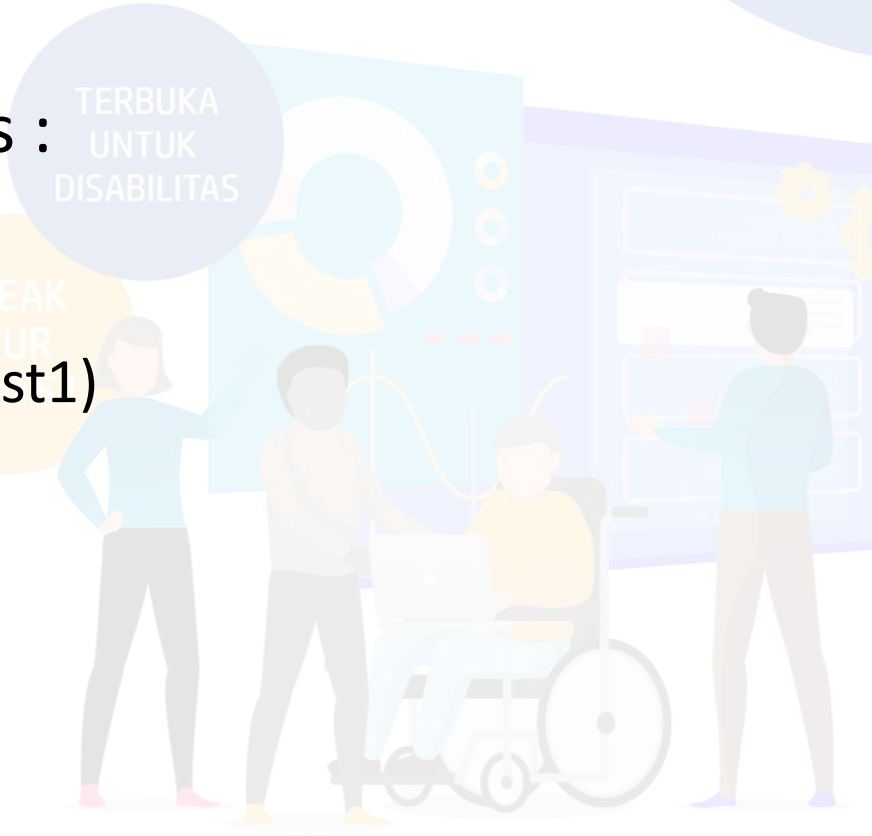
- Amazon EC2 provides you with flexible, cost effective, and easy-to-use data storage options for your instances.
- Each option has a unique combination of performance and durability. These storage options can be used independently or in combination to suit your requirements.
- After reading this section, you should have a good understanding about how you can use the data storage options supported by Amazon EC2 to meet your specific requirements.
- These storage options include the following:
  - Amazon Elastic Block Storage (EBS)
  - Amazon EC2 Instance Store
  - Amazon Elastic File System (EFS)
  - Amazon Simple Storage Service (S3)

Elastic Block Storage

- Amazon Elastic Block Store (Amazon EBS) provides block level storage volumes for use with EC2 instances.
- EBS volumes behave like raw, unformatted block devices. You can mount these volumes as devices on your instances.
- EBS volumes that are attached to an instance are exposed as storage volumes that persist independently from the life of the instance.
- You can create a file system on top of these volumes, or use them in any way you would use a block device (such as a hard drive).

EBS Types

- EBS provides following types :
  - General Purpose SSD (gp2)
  - Provisioned IOPS SSD (io1)
  - Throughput Optimized HDD (st1)
  - Cold HDD (sc1)

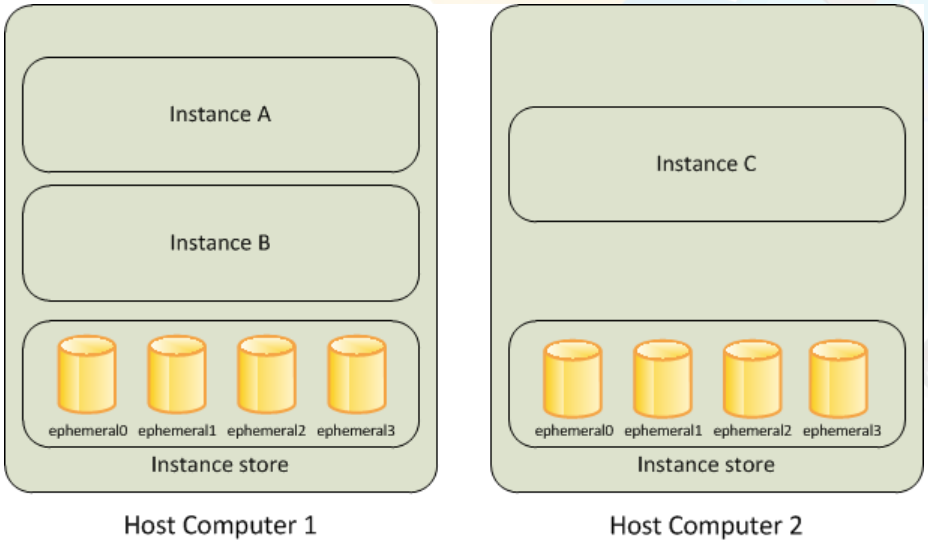


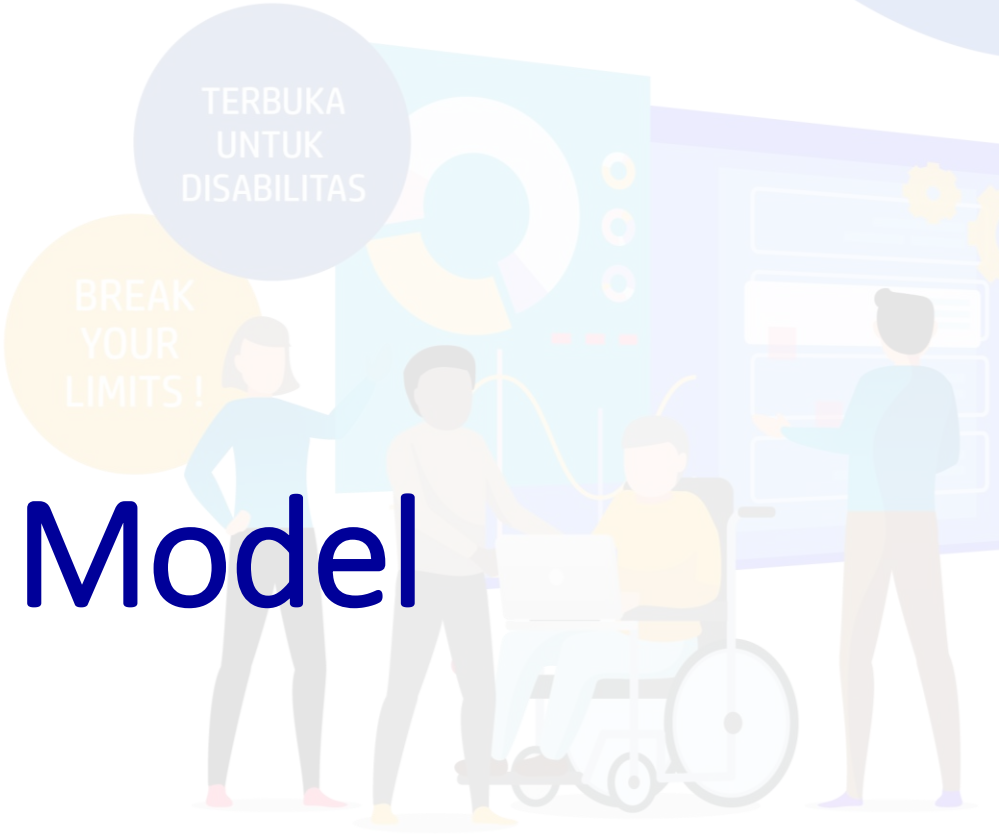
Instance Store

- An *instance store* provides temporary block-level storage for your instance.
- This storage is located on disks that are physically attached to the host computer.
- Instance store is ideal for temporary storage of information that changes frequently, such as buffers, caches, scratch data, and other temporary content, or for data that is replicated across a fleet of instances, such as a load-balanced pool of web servers.

Instance Store (Cont'd)

- An instance store consists of one or more instance store volumes exposed as block devices.
- The size of an instance store as well as the number of devices available varies by instance type.





AWS Pricing Model

Amazon EC2 Purchasing Options

On-Demand Instances	Reserved Instances	Scheduled Instances	Spot Instances	Dedicated Instances	Dedicated Hosts
Pay by the **hour**.	Purchase, at a significant **discount**, instances that are **always available.** 1-year to 3-year terms.	Purchase instances that are **always available** on the specified **recurring schedule**, for a one-year term.	Bid on **unused instances**, which can run as long as they are available and your bid is above the Spot price.	Pay, by the hour, for instances that run on **single-tenant hardware**.	Pay for a physical host that is **fully dedicated** to running your instances.

Pricing Model

AWS offers several pricing models depending on product. These include:

- **On Demand** means you pay for compute or database capacity with no long-term commitments or upfront payments.
- **Dedicated Instances** (available with Amazon Elastic Compute Cloud (Amazon EC2)) run in a virtual private cloud (VPC) on hardware that's dedicated to a single customer.
- **Spot Instances** are an Amazon EC2 pricing mechanism that lets you purchase spare computing capacity with no upfront commitment at discounted hourly rates.
- **Reservations** provide you with the ability to receive a greater discount, up to 75 percent, by paying for capacity ahead of time. More detail is provided in the section, "Optimizing costs with reservations

Case Study

Case Study : Qlue

- Qlue is an Indonesian AI and IoT-based smart city ecosystem company. It implements technology for a sustainable, safer city, and works closely with city institutions, local communities, and police departments to solve urban issues related to mobility, security & safety, and sanitation.
- Leveraging AWS, it developed QlueWork, a mobile app for managing and coordinating in-field tasks, and QlueDashboard, whose data integration assists real-time monitoring, to help improve efficiency and effectiveness on resource management and allocation.
- Qlue has been able to reduce time to market by 50% with streamlined deployment process, with the ability to scale seamlessly while processing large volumes of real time data.

Key Point

- AWS provides many services including computing, storage, and networking
- AWS provides wide range service for computing
- The most common computing service is EC2
- AWS provides wide range of EC2 specifications, called Instance Types
- AWS provides wide range of storage types that can be attached to EC2 instances
- AWS provides multiple pricing scheme that can be chosen to fit customer's workload

Follow our social media!

DIGITAL TALENT SCHOLARSHIP

- digitalent.kominfo
- digitalent.kominfo
- DTS_kominfo
- Digital Talent Scholarship 2020

Pusat Pengembangan Profesi dan Sertifikasi
Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika
Jl. Medan Merdeka Barat No. 9
(Gd. Belakang Lt. 4 - 5)
Jakarta Pusat, 10110

digitalent.kominfo.go.id