EC2

- This information is stored in the instance metadata on the instance. You can access the instance metadata through a URI or by using the Instance Metadata Query tool.

- The instance metadata is available at http://169.254.169.254/latest/meta-data.

- The Instance Metadata Query tool allows you to query the instance metadata without having to type out the full URI or category names.

- You can attach a network interface to an instance when it's running (hot attach), when it's stopped (warm attach), or when the instance is being launched (cold attach).

- By default, AWS has a limit of 20 instances per region. This includes all instances set up on your AWS account.
  To increase EC2 limits, request a higher limit by providing information about the new limit and regions where it should be applied.

- Amazon EC2 uses public key cryptography to encrypt and decrypt login information.

- Public key cryptography uses a public key to encrypt a piece of data, and then the recipient uses the private key to decrypt the data. The public and private keys are known as a key pair.

- Public key cryptography enables you to securely access your instances using a private key instead of a password.

- A key pair consists of a public key that AWS stores, and a private key file that you store:
  – For Windows AMIs, the private key file is required to obtain the password used to log into your instance.
  – For Linux AMIs, the private key file allows you to securely SSH into your instance.

when u restart ec2 instance and it immediately changed from a pending state to a terminated state.

- The following are a few reasons why an instance might immediately terminate:
  – You've reached your EBS volume limit.
  – An EBS snapshot is corrupt.
  – The root EBS volume is encrypted and you do not have permissions to access the KMS key for decryption.
  – The instance store-backed AMI that you used to launch the instance is missing a required part (an image.part.xx file).

- Instances in standby state are still managed by Auto Scaling, are charged as normal, and do not count towards available EC2 instance for workload/application use. Auto scaling does not perform health checks on instances in the standby state. Standby state can be used for performing updates/changes/troubleshooting etc. without health checks being performed or replacement instances being launched.

EBS volumes

- An Amazon EBS volume is a durable, block-level storage device that you can attach to your instances.
- After you attach a volume to an instance, you can use it as you would use a physical hard drive.
- EBS volumes are flexible.
- When you create an EBS volume, it is automatically replicated within its Availability Zone to prevent data loss due to the failure of any single hardware component.
- You can attach an EBS volume to an EC2 instance in the same Availability Zone.
- non-root EBS volumes remain available even after you terminate an instance to which the volumes were attached
- When you launch an instance, the root device volume contains the image used to boot the instance. You can choose between AMIs backed by Amazon EC2 instance store and AMIs backed by Amazon EBS.
- By default, the root volume for an AMI backed by Amazon EBS is deleted when the instance terminates. You can change the default behavior to ensure that the volume persists after the instance terminates.

EBS Volume Types

- **General Purpose SSD (gp2) Volumes**
  - They were designed to be a cost-effective storage option for a wide variety of workloads.
  - Gp2 volumes cover system volumes, dev and test environments, and various low-latency apps. They come in sizes between 1Gib and 16Gib and provide very low latency, down to single-digit milliseconds.
  - They have a decent IOPS (starting from 100 and going all the way to 16000 IOPS) and a maximum throughput of 250MiB/s.
  - You can combine multiple EBS volume types in a RAID to achieve even higher performance on a single instance.
  - Gp2 volumes are fairly cheap, especially for the balanced performance they provide. They are priced at $0.1 per GB per month of provisioned storage.
- **Provisioned IOPS SSD (io1) Volumes**
  - Provisioned IOPS SSD (io1) EBS volume types are a special type of volume created to fulfill the needs of very intensive I/O workloads that require very high throughput.
  - They are useful for cases which are latency-sensitive, like large database workloads (e.g., MySQL, Cassandra, MongoDB, and Oracle) and critical business applications that need the kind of sustained performance gp2 volumes can't achieve.
  - Io1 volumes can store between 4GiB and 16TiB, and their IOPS can be as low as 100 or high as 64000 IOPS per volume with up to 1,000 MiB/s of throughput.
  - While the performance of gp2 volumes is dictated by volume size, the performance of provisioned IOPS SSD volumes can be set during creation time. It is limited by a maximum IOPS to volume size ratio of 50:1.
  - Of course, all of this comes with a price—$0.125 per GB per month of provisioned storage and $0.065 per provisioned IOPS per month. This can amount to a costly solution in some cases, so be careful when provisioning these volumes.
- **Throughput Optimized HDD (st1) Volumes**
  - Throughput Optimized HDD (st1) volumes are a type of volume that offers low-cost storage while fulfilling the need for sequential workloads that require more throughput than IOPS.
  - When working with data warehouses, log processing, ETL (extract, transform, load) or AWS EMR, this is a volume type to look into.
  - Keep in mind that this volume type cannot be used as a boot volume. st1, like gp2, relies on burstable performance, and volume size will be the main factor when calculating baseline performance.
  - St1 volumes can range in size between 500GiB and 16TiB, and they allow for 500 MiB/s of throughput per volume with 500 IOPS.

- St1 volumes also come with a lower price tag: $0.045 per GB per month of provisioned storage.

- **Cold HDD (sc1) Volumes**
  - Cold HDD (sc1) volumes, like st1 volumes, provide low-cost storage for workloads that rely on throughput rather than IOPS.
  - Sc1 volumes are primarily used for large amounts of data that is infrequently accessed, or in cases where the cost of storage is the most important factor.
  - Sc1, just like st1, cannot be used as a boot volume and relies on burstable performance.
  - Sc1 volumes come in sizes between 500GiB and 16TiB and provide up to 250 IOPS and 250MiB/s of throughput per volume.
  - Sc1 volumes are the cheapest option available, costing only $0.025 per GB per month of

EBS – Volume Types Summary

	General Purpose SSD		Provisioned IOPS SSD		
Volume type	gp3	gp2	io2 Block Express ‡	io2	io1
Durability	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	99.999% durability (0.001% annual failure rate)		99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)
Use cases	• Low-latency interactive apps • Development and test environments		Workloads that require sub-millisecond latency, and sustained IOPS performance or more than 64,000 IOPS or 1,000 MiB/s of throughput	• Workloads that require sustained IOPS performance or more than 16,000 IOPS • I/O-intensive database workloads	
Volume size	1 GiB - 16 TiB		4 GiB - 64 TiB	4 GiB - 16 TiB	
Max IOPS per volume (16 KiB I/O)	16,000		256,000	64,000 †	

	Throughput Optimized HDD	Cold HDD
Volume type	st1	sc1
Durability	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)
Use cases	• Big data • Data warehouses • Log processing	• Throughput-oriented storage for data that is infrequently accessed • Scenarios where the lowest storage cost is important
Volume size	125 GiB - 16 TiB	125 GiB - 16 TiB
Max IOPS per volume (1 MiB I/O)	500	250
Max throughput per volume	500 MiB/s	250 MiB/s
Amazon EBS Multi-attach	Not supported	Not supported
Boot volume	Not supported	Not supported

EBS resiliency (across Multiple AZ)

- You can back up the data on your Amazon EBS volumes to Amazon S3 by taking point-in-time snapshots.
- Snapshots are incremental backups, which means that only the blocks on the device that have changed after your most recent snapshot are saved.
- This minimizes the time required to create the snapshot and saves on storage costs by not duplicating data.
- When you delete a snapshot, only the data unique to that snapshot is removed.

- Each snapshot contains all of the information that is needed to restore your data (from the moment when the snapshot was taken) to a new EBS volume.

EFS (Elastic file System)

- With EFS you can transition files to EFS IA after a file has not been accessed for a specified period of time with options up to 90 days.

Instance Store

- You can specify instance store volumes for an instance only when you launch it. You can't detach an instance store volume from one instance and attach it to a different instance.
- The data in an instance store persists only during the lifetime of its associated instance. If an instance reboots (intentionally or unintentionally), data in the instance store persists.
- If you create an AMI from an instance, the data on its instance store volumes isn't preserved and isn't present on the instance store volumes of the instances that you launch from the AMI.
- When you stop, hibernate, or terminate an instance, every block of storage in the instance store is reset. Therefore, this option is incorrect.
- You can specify instance store volumes for an instance only when you launch it.
- 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.

Placement Group

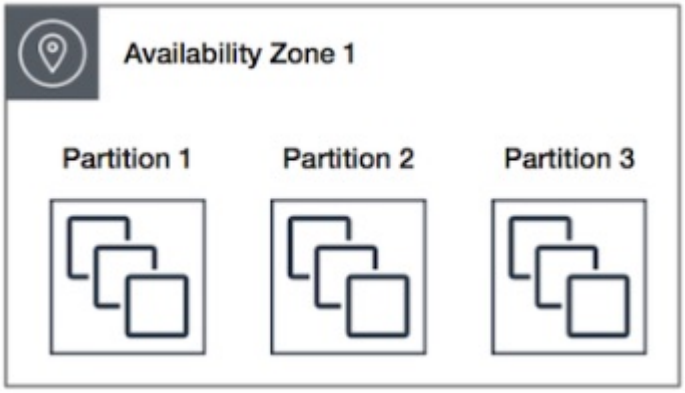
Partition placement groups - Partition placement groups help reduce the likelihood of correlated hardware failures for your application. When using partition placement groups, Amazon EC2 divides each group into logical segments called partitions. Amazon EC2 ensures that each partition within a placement group has its own set of racks. Each rack has its own network and power source. No two partitions within a placement group share the same racks, allowing you to isolate the impact of a hardware failure within your application.

The following image is a simple visual representation of a partition placement group in a single Availability Zone. It shows instances that are placed into a partition placement group with three partitions—Partition 1, Partition 2, and Partition 3. Each partition comprises multiple instances. The instances in a partition do not share racks with the instances in the other partitions, allowing you to contain the impact of a single hardware failure to only the associated partition.

Partition placement groups can be used to deploy large distributed and replicated workloads, such as HDFS, HBase, and Cassandra, across distinct racks. When you launch instances into a partition placement group, Amazon EC2 tries to distribute the instances evenly across the

number of partitions that you specify. You can also launch instances into a specific partition to have more control over where the instances are placed.

A partition placement group can have partitions in multiple Availability Zones in the same Region. A partition placement group can have a maximum of seven partitions per Availability Zone. The number of instances that can be launched into a partition placement group is limited only by the limits of your account.



Cluster placement groups - A cluster placement group is a logical grouping of instances within a single Availability Zone. A cluster placement group can span peered VPCs in the same Region. Instances in the same cluster placement group enjoy a higher per-flow throughput limit for TCP/IP traffic and are placed in the same high-bisection bandwidth segment of the network. Cluster placement groups are recommended for applications that benefit from low network latency, high network throughput, or both. They are also recommended when the majority of the network traffic is between the instances in the group. As the instances are packed close together inside an Availability Zone, this option is not correct for the given use case.

Spread placement groups - A spread placement group is a group of instances that are each placed on distinct racks, with each rack having its own network and power source. Spread placement groups are recommended for applications that have a small number of critical instances that should be kept separate from each other. Launching instances in a spread placement group reduces the risk of simultaneous failures that might occur when instances share the same racks. Spread placement groups provide access to distinct racks, and are

therefore suitable for mixing instance types or launching instances over time. As the use-case talks about running large distributed and replicated workloads, so it needs more instances, therefore this option is not the right fit for the given use-case.

A spread placement group can span multiple Availability Zones in the same Region. You can have a maximum of seven running instances per Availability Zone per group.

The following image shows seven instances in a single Availability Zone that are placed into a spread placement group. The seven instances are placed on seven different racks.