



EC2

User Data

É possível incluir um script que rodará na primeira vez que a instância for iniciada, chamado user data, neste script é possível automatizar tarefas como:

- Instalação de atualizações
- Instalação de softwares
- Baixar arquivos da internet
- E muitas outras coisas...

O script de User Data roda na conta do usuário root, por tanto todos os comandos dados tem o privilégios sudo.

EC2 instance types: example

Instance	vCPU	Mem (GiB)	Storage	Network Performance	EBS Bandwidth (Mbps)
t2.micro	1	1	EBS-Only	Low to Moderate	
t2.xlarge	4	16	EBS-Only	Moderate	
c5d.4xlarge	16	32	1 x 400 NVMe SSD	Up to 10 Gbps	4,750
r5.16xlarge	64	512	EBS Only	20 Gbps	13,600
m5.8xlarge	32	128	EBS Only	10 Gbps	6,800

Detalhes das Instâncias

O Nome

- AWS has the following naming convention:

m5.2xlarge

- m: instance class
- 5: generation (AWS improves them over time)
- 2xlarge: size within the instance class

Os Tipos

• General Purpose:

- Equilíbrio entre processamento, memória e rede.
- Estarei utilizando o t2.micro, que é uma instância do tipo *General Purpose*.
- **Use case:** Sites e aplicativos web, ambientes de desenvolvimento, servidores de compilação repositórios de código, microsserviços, ambientes de teste e preparação e aplicativos de linha de negócios.

Uso geral

Instâncias de uso geral fornecem um equilíbrio de recursos de computação, memória e rede e podem ser usadas para diversas cargas de trabalho. Essas instâncias são ideais para aplicativos que usam esses recursos em proporções iguais, como servidores web e repositórios de código.

M7g	M7i	M7i-flex	M7a	Mac	M6g	M6i	M6in	M6a	M5	M5n	M5zn	M5a
M4	T4g	T3	T3a	T2								

As [instâncias T2 do Amazon EC2](#) são instâncias expansíveis que oferecem um nível de referência de performance de CPU com capacidade de expansão acima da referência.

As instâncias T2 ilimitadas podem manter a alta performance de CPU pelo tempo que a workload precisar. Para a maioria das cargas de trabalho de uso geral, as instâncias de T2 ilimitado fornecerão performance ampla sem cobranças adicionais. Se for necessário executar a instância com uma utilização maior da CPU para um período de tempo prolongado, isso também poderá ser feito com uma tarifa contínua de 5 centavos por hora de vCPU.

A performance de referência e a capacidade de intermitência são regidas pelos créditos de CPU. As instâncias T2 recebem continuamente créditos de CPU a uma determinada razão, dependendo do tamanho da instância, e acumulam créditos de CPU quando ociosas e consomem esses créditos quando ativas. As instâncias T2 são uma boa opção para diversas cargas de trabalho de uso geral, incluindo microsserviços; aplicativos interativos de baixa latência; bancos de dados de pequeno e médio portes; desktops virtuais; ambientes de desenvolvimento, compilação e preparação; repositórios de código e protótipos de produtos. Para mais informações, leia [Instâncias de performance com capacidade de intermitência](#).

- **Compute Optimized:**

- Instância que possui foco no poder de processamento (CPU).
- **Use case:** Computação de alta performance (HPC), processamento em lote, veiculação de anúncios, codificação de vídeo, jogos, modelagem científica, análise distribuída e inferência de machine learning com base em CPU.
-

Otimizadas para computação

As instâncias otimizadas para computação são ideais para aplicativos vinculados a computação que se beneficiam de processadores de alta performance. As instâncias pertencentes a essa categoria são adequadas para workloads de processamento em lote, transcodificação de mídia, servidores da web de alta performance, computação de alta performance (HPC), modelagem científica, servidores de jogos dedicados e mecanismos de servidor de anúncios, inferência de machine learning e outras aplicações com uso intensivo de computação.

C7g	C7gn	C7i	C7i-flex	C7a	C6g	C6gn	C6i	C6in	C6a	C5	C5n	C5a
C4												
As instâncias C7g do Amazon EC2 são alimentadas por processadores AWS Graviton3 baseados em Arm. Oferecem a melhor relação entre preço e performance no Amazon EC2 para aplicações com uso intensivo de computação.												

- **Memory Optimized:**

- São projetadas para fornecer performance rápida para workloads que processam grandes conjuntos de dados na memória.
- **Use case:** Workloads que consomem muita memória, como bancos de dados de código aberto, caches na memória e análise de big data em tempo real.

Otimizadas para memória

As instâncias otimizadas para memória são projetadas para fornecer performance rápida para workloads que processam grandes conjuntos de dados na memória.

R8g	R7g	R7i	R7iz	R7a	R6g	R6i	R6in	R6a	R5	R5n	R5b	R5a	R4
U7i	Mais memória (U-1)		X2gd	X2idn	X2iedn	X2iezn	X1	X1e	z1d				

As [instâncias R8g do Amazon EC2](#) utilizam tecnologia de processadores AWS Graviton4. Elas proporcionam a melhor relação entre custo e desempenho no Amazon EC2 para aplicações que requerem otimização de memória.

- **Storage Optimized:**

- Ótimas para tarefas que exigem alta capacidade de armazenamento, isto é, acesso a leitura e escrita de uma grande quantidade de dados no disco local.
- **Use case:** Aplicações intensivas de E/S, são instâncias direcionadas a clientes que usam bancos de dados transacionais (Amazon DynamoDB, MySQL e PostgreSQL), Amazon OpenSearch Service, Aplicações de Data warehouse e análises em tempo real, como o Apache Spark.

Otimizadas para armazenamento

As instâncias otimizadas para armazenamento são projetadas para cargas de trabalho que exigem acesso de leitura e gravação sequencial alto a conjuntos de dados muito grandes no armazenamento local. São otimizadas para fornecer dezenas de milhares de operações de E/S aleatórias de baixa latência por segundo (IOPS) para aplicações.

I4g	Im4gn	Is4gen	I4i	I3	I3en	D3	D3en	D2	H1
-----	-------	--------	-----	----	------	----	------	----	----

As [instâncias I4g do Amazon EC2](#) são alimentadas por processadores AWS Graviton2 e oferecem a melhor relação preço-desempenho para workloads com uso intensivo de armazenamento no Amazon EC2. As instâncias I4g oferecem um desempenho computacional até 15% melhor em comparação com instâncias semelhantes otimizadas para armazenamento.

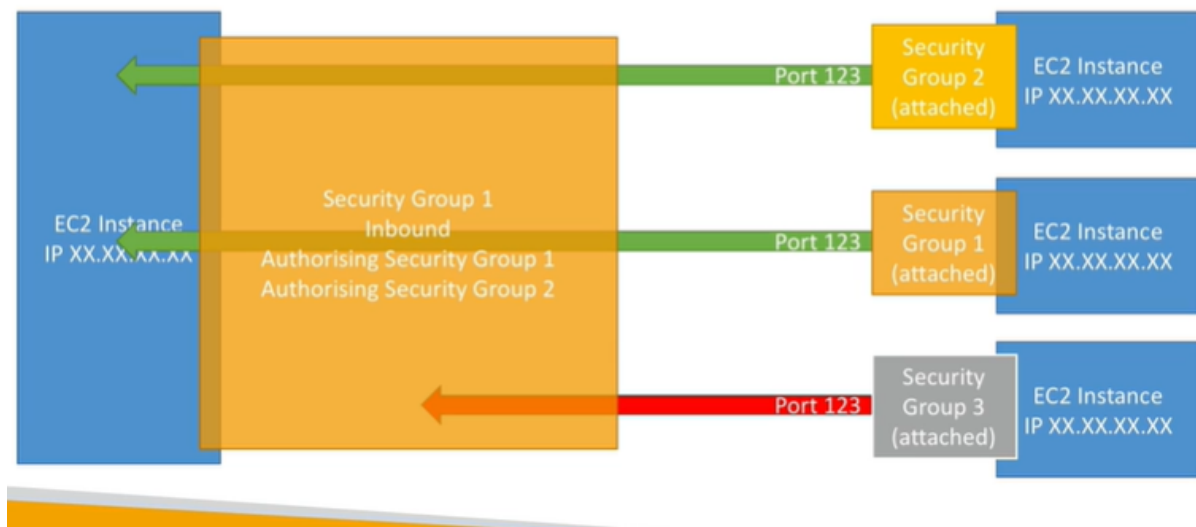
Recursos:

Grupos de Segurança

- Os **Security Groups** são o ponto chave da segurança de rede na AWS.
- Eles controlam como o tráfego de E/S é permitido ou negado para a EC2.
- Security Groups possuem somente regras de allow
- As regras dos **Security Groups** podem referenciar por IP ou por outro Security Group (**SGs podem referenciar uns aos outros**).

No esquema de referência a outro grupo de segurança, indica que o tráfego está liberado para qualquer instância que faça parte daquele grupo referenciado, independente do IP, bom demais.

Referencing other security groups Diagram

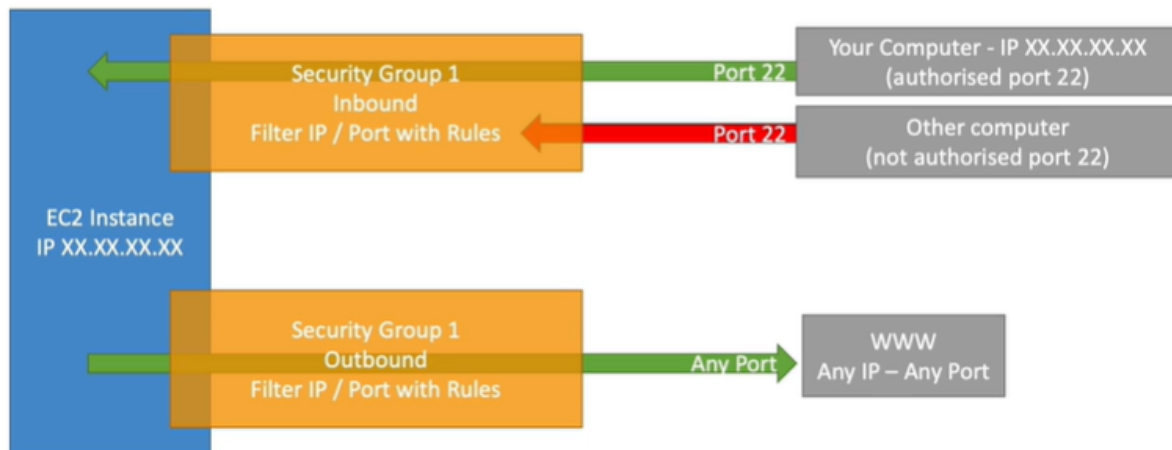


Basicamente, Grupos de Segurança agem como firewalls para instâncias EC2.

Eles regulam:

- Acesso a portas;
- Faixas de IP liberadas (IPv4 e IPv6);
- Controle de conexões **inbound** (de um dispositivo externo para a instância);
- Controle de conexões **outbound** (Da instância para dispositivo externo);

Security Groups Diagram



Opções de Cobrança

- On-Demand Instances – short workload, predictable pricing, pay by second
- Reserved (1 & 3 years)
 - Reserved Instances – long workloads
 - Convertible Reserved Instances – long workloads with flexible instances
- Savings Plans (1 & 3 years) – commitment to an amount of usage, long workload
- Spot Instances – short workloads, cheap, can lose instances (less reliable)
- Dedicated Hosts – book an entire physical server; control instance placement
- Dedicated Instances – no other customers will share your hardware
- Capacity Reservations – reserve capacity in a specific AZ for any duration

Which purchasing option is right for me?



- On demand: coming and staying in resort whenever we like, we pay the full price
- Reserved: like planning ahead and if we plan to stay for a long time, we may get a good discount.
- Savings Plans: pay a certain amount per hour for certain period and stay in any room type (e.g., King, Suite, Sea View, ...)
- Spot instances: the hotel allows people to bid for the empty rooms and the highest bidder keeps the rooms. You can get kicked out at any time
- Dedicated Hosts: We book an entire building of the resort
- Capacity Reservations: you book a room for a period with full price even you don't stay in it

Price Comparison Example – m4.large – us-east-1

Price Type	Price (per hour)
On-Demand	\$0.10
Spot Instance (Spot Price)	\$0.038 - \$0.039 (up to 61% off)
Reserved Instance (1 year)	\$0.062 (No Upfront) - \$0.058 (All Upfront)
Reserved Instance (3 years)	\$0.043 (No Upfront) - \$0.037 (All Upfront)
EC2 Savings Plan (1 year)	\$0.062 (No Upfront) - \$0.058 (All Upfront)
Reserved Convertible Instance (1 year)	\$0.071 (No Upfront) - \$0.066 (All Upfront)
Dedicated Host	On-Demand Price
Dedicated Host Reservation	Up to 70% off
Capacity Reservations	On-Demand Price

Interfaces de Rede

ENI (Elastic Network Interface) - Padrão:

- **Definição:** Uma ENI é uma interface de rede virtual que pode ser anexada a uma instância EC2. Ela funciona como uma interface de rede física, contendo um endereço IP primário, um ou mais endereços IP secundários,

um ou mais grupos de segurança, um endereço MAC, e outras configurações de rede.

- **Uso:** ENIs são usadas para conectar instâncias EC2 a sub-redes em uma VPC. Elas permitem que você mova interfaces entre instâncias, o que pode ser útil para cenários de alta disponibilidade ou para isolar diferentes tipos de tráfego.
- **Exemplo de Aplicação:** Ter múltiplas interfaces de rede em uma instância para segmentar o tráfego de rede ou para separar diferentes camadas de aplicação.

ENA (Elastic Network Adapter) - Alta Performance:

- **Definição:** O ENA é um adaptador de rede de alta performance projetado para oferecer suporte a Enhanced Networking na AWS. Ele oferece maior largura de banda, menor latência, e menor sobrecarga em comparação com as interfaces de rede padrão.
- **Uso:** ENA é usado principalmente em instâncias EC2 que exigem alta performance de rede, como em aplicações de HPC (High-Performance Computing) e grandes clusters de dados.
- **Exemplo de Aplicação:** Quando uma instância EC2 precisa de uma largura de banda de rede de até 100 Gbps, como em simulações científicas ou análises de Big Data.

EFA (Elastic Fabric Adapter) - Potência Máxima:

- **Definição:** O EFA é um adaptador de rede que fornece latência ultrabaixa e alta taxa de transferência, além de suporte para aplicativos de HPC que usam bibliotecas de comunicação de rede como MPI (Message Passing Interface). **O EFA não é compatível com as instâncias do Windows.**
- **Uso:** EFA é especificamente projetado para workloads HPC que requerem uma comunicação de rede extremamente rápida e eficiente, como em simulações científicas, modelagem financeira ou análises de petróleo e gás.
- **Exemplo de Aplicação:** Aplicações HPC distribuídas que exigem uma comunicação rápida entre instâncias EC2 em um cluster, como no processamento de fluido dinâmico computacional (CFD) ou em renderização gráfica em 3D.

Web Application Firewall (WAF)

Com este serviço, é possível monitorar as requisições HTTP/HTTPS que chegam em seus servidores, também é possível controlar quem acessa os seus conteúdos. É possível fazer diversas condições de bloqueio para requisições, por exemplo:

- Endereço de IP da origem
- País de onde a requisição se originou
- Valores na header
- Strings que aparecem no request body (regex)
- Tamanho da requisição
- Presença de código SQL (famoso SQL Injection)
- Presença de script malicioso (Cross-site scripting)

Placement Groups

São um recurso do Amazon EC2 que permite controlar o posicionamento de instâncias em hardware subjacente dentro de uma Availability Zone da AWS. Eles são usados para melhorar o desempenho da rede ou a resiliência das instâncias.

Tipos de Placement Groups:

1. Cluster:

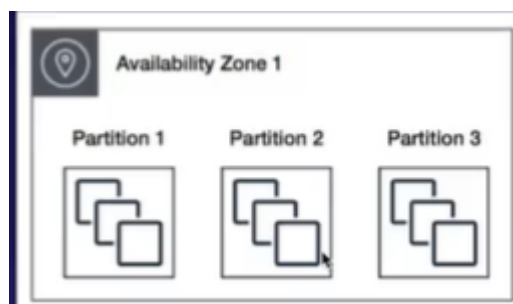
- **Objetivo:** Maximizar o desempenho da rede ao agrupar instâncias próximas umas das outras no mesmo rack ou em racks adjacentes.
- **Uso Ideal:** Workloads que exigem baixa latência e alta taxa de transferência entre instâncias, como HPC (computação de alto desempenho) e grandes bancos de dados distribuídos.
- **Limitação:** Alta disponibilidade pode ser comprometida, pois as instâncias estão fisicamente próximas e podem ser afetadas por falhas no rack.

2. Spread:

- **Objetivo:** Garantir que as instâncias sejam distribuídas por diferentes racks, reduzindo o risco de falhas correlacionadas.
- **Uso Ideal:** Aplicações críticas que precisam de alta disponibilidade e resiliência, onde as instâncias precisam estar em racks diferentes.
- **Limitação:** Número máximo de instâncias por grupo espalhado é limitado (normalmente até 7 instâncias por grupo em algumas regiões).

3. Partition:

- **Objetivo:** Distribuir instâncias em partições lógicas, onde cada partição tem seus próprios racks. Isso fornece isolamento entre grupos de instâncias dentro de uma partição. Partições podem estar em múltiplas AZ.
- **Uso Ideal:** Grandes sistemas distribuídos, como Hadoop, HDFS, e Cassandra, onde a falha de uma partição não deve afetar as outras.
- **Limitação:** Menor densidade de instâncias por partição comparado ao Cluster Placement Group.
- A diferença entre Partition e Spread é que uma partição que fica em um rack pode incluir várias instancias, veja a imagem de exemplo onde temos 3 partições, cada uma em sua própria rack, e em cada partição temos 3 instâncias:



Cenários de Uso:

- **Cluster Placement Group:** Quando você precisa de baixa latência de rede, como em simulações científicas, machine learning, ou grandes bancos de dados.
- **Spread Placement Group:** Quando a prioridade é a resiliência, como em sistemas críticos que não podem ter todas as instâncias afetadas por uma


única falha de hardware.

- **Partition Placement Group:** Para grandes sistemas distribuídos que requerem isolamento de falhas entre diferentes grupos de dados ou componentes do sistema.

Resumo:

- **Cluster Placement Groups** são usados para otimizar o desempenho da rede.
- **Spread Placement Groups** garantem a alta disponibilidade distribuindo instâncias em racks diferentes.
- **Partition Placement Groups** oferecem uma abordagem híbrida que isola falhas entre partições.

Outros Detalhes

- Um único SG pode estar atrelado a múltiplas instâncias
- SGs são bloqueados a uma combinação região / VPC
- O usuário da EC2 não visualiza as regras de tráfego aplicadas a máquina dele
- É uma boa prática manter um SG exclusivo para acesso SSH.
- **Troubleshooting:**
 - Se a sua aplicação não está acessível (timeout), é problema no SG.
 - Agora, se a aplicação devolve um erro "connection refused", então é um erro da aplicação, ou então ela não está aberta mesmo, pois o tráfego está ok.
- Por padrão:
 - **Todo tráfego inbound é bloqueado.**
 - **Todo tráfego outbound está autorizado.**
- Para informações mais específicas, confira o módulo 6 das anotações do Osvalí.
 -  Módulo 6: Computação

-  Spot Fleets

Porta para lembrar:

- 3389 ⇒ RDP (Remote Desktop Protocol) - Logar em uma instância Windows.

Avançado

- Elastic IP é o nome dado a opção de IP estático para as instâncias EC2.

 EBS