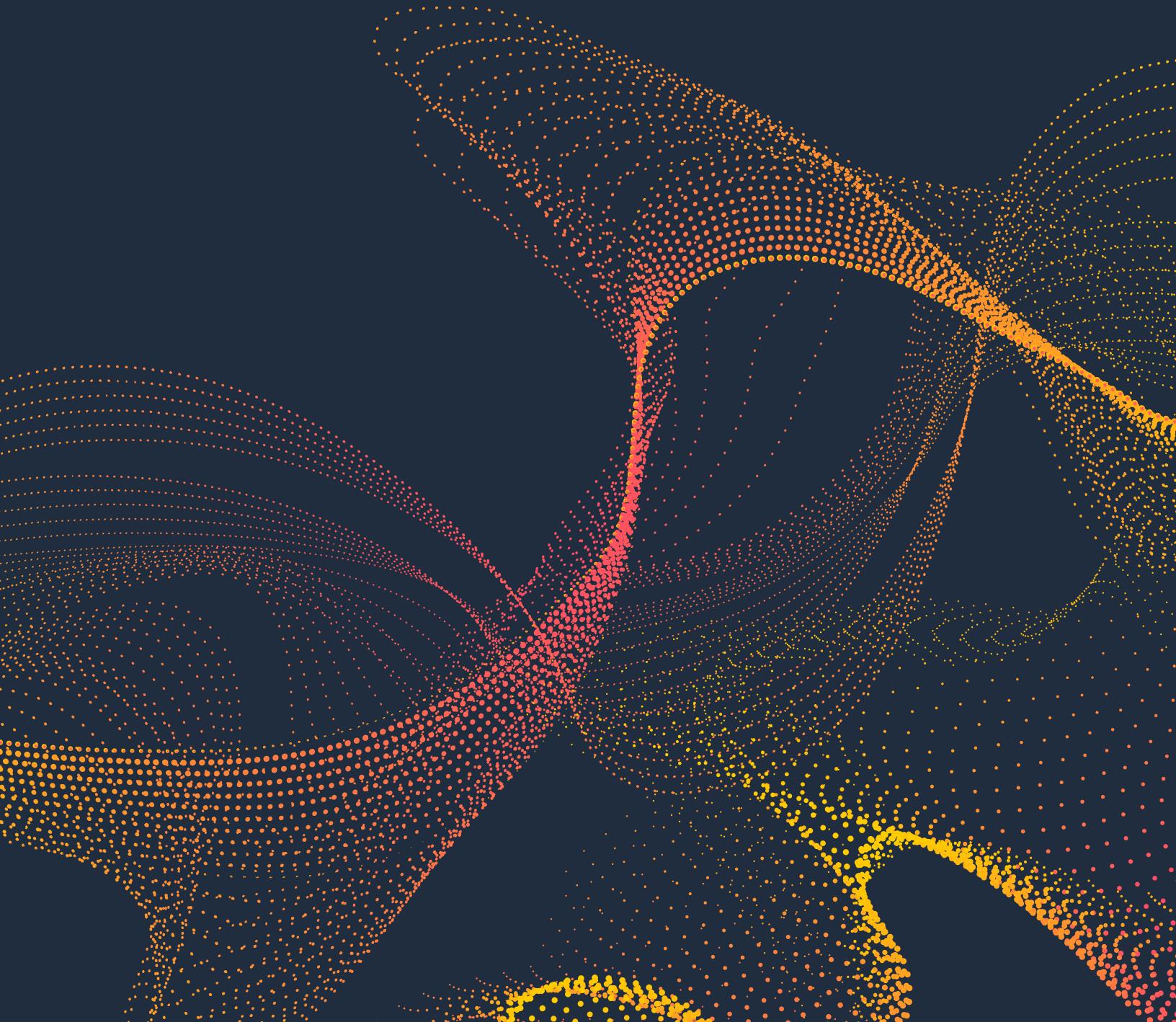


# O custo total de propriedade (TCO) do Amazon SageMaker



# TCO do Amazon SageMaker

O [Amazon SageMaker](#) é um serviço modular de machine learning (ML) de ponta a ponta para criar, treinar e implementar modelos em escala e com custos significativamente mais baixos que outras opções de ML. O custo total de propriedade (Total Cost of ownership, TCO) do Amazon SageMaker em um período de três anos é 54% mais baixo em comparação com outras opções de ML baseadas na nuvem, como as opções [Amazon EC2](#) (EC2) autogerenciadas e o [Elastic Kubernetes Service](#) (EKS) da AWS. Além do TCO mais baixo, os recursos totalmente gerenciados e integrados do Amazon SageMaker permitem colocar as ideias de ML em produção mais rapidamente e melhorar a produtividade do cientista de dados em até 10 vezes.



A Coinbase usa modelos de ML no Amazon SageMaker para ajudar na prevenção de fraudes, verificação de identidade e conformidade em larga escala. **O uso do Amazon SageMaker reduziu o tempo de treinamento do modelo de 20 horas para 10 minutos.**



A Intuit desenvolveu modelos de ML que podem analisar um ano de transações bancárias para encontrar despesas comerciais dedutíveis para os clientes. **Usando o Amazon SageMaker, o Intuit reduziu o tempo de implantação do ML em 90%, de 6 meses para 1 semana.**

Equipes de todos os tamanhos podem se beneficiar de um TCO significativamente menor ao usar o Amazon SageMaker. Por exemplo, em três anos, uma equipe pequena de cinco cientistas de dados pode esperar um TCO 96% menor usando o Amazon SageMaker em comparação à criação e manutenção de sua própria plataforma de ML no EC2 ou EKS. Equipes de tamanho médio com 15 cientistas de dados podem esperar um TCO 90% menor com o Amazon SageMaker. Equipes grandes com 50 cientistas de dados podem esperar um TCO 78% menor com o Amazon SageMaker comparado ao EC2 e 65% menor comparado ao EKS na AWS. Equipes ainda maiores com 250 cientistas de dados podem esperar um TCO 72% menor com o Amazon SageMaker em comparação com o EC2 e 54% menor em comparação com o EKS na AWS.

## Equipes que usam o Amazon SageMaker

RESUMO GERAL	Economia de TCO em três anos no Amazon SageMaker	
	Em comparação com o EC2	Em comparação com o EKS
Cenário pequeno	-96%	-96%
Cenário médio	-91%	-90%
Cenário grande	-78%	-65%
Cenário muito grande	-72%	-54%



# O cálculo do TCO do Amazon SageMaker incluiu a avaliação de custos para cada uma das três fases do ML:

## Criar

Explore, faça pré-processamento de dados e experimente estruturas e algoritmos de ML usando blocos de anotações

## Treinar

Treine modelos de ML e ajuste seus hiperparâmetros em grandes conjuntos de dados

## Implantar

Coloque modelos de ML em produção e faça inferências sobre dados invisíveis

Cada uma dessas fases incluía o custo de execução da infraestrutura (computação, armazenamento e rede), os custos operacionais e os custos de criação de segurança e conformidade. Os custos de computação, armazenamento e infraestrutura de rede são medidos com base no uso. Os custos operacionais para operação, monitoramento e manutenção da infraestrutura são medidos com base no salário dos engenheiros. As cargas de trabalho de ML são inherentemente dependentes de dados em todo o ciclo de vida. As organizações assumem gastos para proteger as cargas de trabalho de ML, obter conformidade com os padrões regulamentares e manter a segurança e a conformidade. Os custos de segurança e conformidade são medidos como o custo dos recursos de engenharia e respectivos salários.

Normalmente, o custo total de propriedade do Amazon SageMaker é menor no primeiro ano em comparação com as opções EC2 ou EKS devido a investimentos mais altos na criação de recursos de segurança e conformidade necessários. Esses recursos vêm prontos para uso no Amazon SageMaker, mas precisam ser criados nessas outras plataformas. O custo total de propriedade do Amazon SageMaker continua a ser significativamente menor do que outras plataformas nos anos subsequentes. Isso se deve à utilização muito maior que o Amazon SageMaker alcança em comparação ao EC2 ou EKS, além dos custos contínuos de manutenção da infraestrutura, segurança e conformidade principais ao escolher as opções EC2 ou EKS autogerenciadas. No Amazon SageMaker, todas elas são fornecidas e estão prontas para uso.

# Serviços considerados para a análise de TCO



A análise de TCO nesta publicação considerou as seguintes opções:

1

**Autogerenciada no Amazon EC2** – Essa opção oferece a você uma abordagem “faça você mesmo” para o ML usando os serviços da AWS como [AWS Batch](#) e [Amazon ECS](#). Você assume a responsabilidade de provisionar e gerenciar instâncias do Amazon EC2, incluindo recuperação de falhas de instância, aplicação de patches e escalabilidade automática. Você usa o DLAMI com as estruturas e bibliotecas do ML pré-criadas, mas precisa otimizar o acesso aos dados para obter alto throughput, otimizar sua configuração para escala e habilitar o treinamento distribuído. Além disso, você precisa criar e manter os recursos de segurança e conformidade necessários para suas cargas de trabalho de ML.

2

**Kubernetes gerenciados na AWS** – Serviços como o [Amazon EKS](#) facilitam a implantação, o gerenciamento e a escalabilidade de cargas de trabalho conteinerizadas no Amazon EC2. No entanto, você precisa assumir o custo adicional de gerenciar seu próprio cluster e ajustar a performance e a utilização com base nos requisitos de memória, computação e rede para suas cargas de trabalho. Além disso, você precisa criar o nível certo de segurança, conformidade e disponibilidade para suas cargas de trabalho de ML. Embora existam ferramentas de código aberto, como o Kubeflow, que facilitam a execução das cargas de trabalho de ML, você ainda assume custos de gerenciamento de infraestrutura, pois o conhecimento necessário para criar, gerenciar e proteger um cluster Kubernetes é muito mais alto e menos disponível.

3

**Serviço de ML totalmente gerenciado** - O [Amazon SageMaker](#) é totalmente gerenciado para que você não precise criar, gerenciar ou manter nenhuma infraestrutura ou ferramenta para criar, treinar ou implantar modelos de ML. Você não precisa gerenciar nenhuma instância, aplicação de patches, atualização ou mesmo criar recursos de escalabilidade automática ou balanceamento de carga para escalar suas cargas de trabalho. Isso reduz o suporte operacional necessário para menos de um décimo em comparação com as soluções autogerenciadas. Você pode selecionar instâncias otimizadas de computação, memória ou gráficos e alcançar uma alta utilização da instância. Além disso, você paga pelo armazenamento e pela rede com base no seu uso. O Amazon SageMaker conta com segurança e conformidade integradas para cargas de trabalho de ML e não exige que você invista em segurança adicional para suas cargas de trabalho.

# Custos de segurança e conformidade



As cargas de trabalho de ML dependem de grandes volumes de dados, muitos dos quais são confidenciais ou proprietários, para criar e treinar modelos. Quando os modelos estão em produção, os modelos de ML precisam garantir solicitações de inferência e respostas.

Embora as plataformas subjacentes ao EC2 e EKS possibilitem segurança e conformidade, você ainda precisa criar a segurança e a conformidade para a configuração do ML, o que gera custos adicionais de engenharia ao usar soluções EC2 ou EKS autogerenciadas. Isso ocorre porque você precisa proteger dados, controlar o acesso e manter a conformidade ao usar blocos de anotações, treinar vários nós e executar inferências nos dados do cliente. Por outro lado, o Amazon SageMaker foi projetado para executar aplicativos de ML em escala com a melhor segurança, conformidade e disponibilidade mundial, para que você não precise criar esses recursos.

## Custos de segurança

O Amazon SageMaker é totalmente seguro com criptografia de ponta a ponta em repouso e em trânsito, incluindo criptografia do volume raiz e do volume EBS, suporte ao [Amazon VPC](#), [AWS Private Link](#), chaves gerenciadas pelo cliente, controle de acesso refinado do [IAM](#), auditorias do [AWS CloudTrail](#), criptografia entre nós para treinamento, controle de acesso baseado em tags, isolamento de rede e Interactive Application Proxy. O Amazon SageMaker economiza mais de 89 meses de esforço de engenharia para negócios grandes e muito grandes, o que equivale a 926 mil USD em custos durante um período de três anos. Isso acontece pois ele fornece recursos de segurança prontos para uso, enquanto nas soluções autogerenciadas no EC2 e EKS os mesmos precisam ser criados. Para fins de modelagem, supõe-se que os requisitos de segurança para pequenas e médias empresas sejam mais baixos, necessitando de aproximadamente 25% e 50% dos recursos fornecidos pelo Amazon SageMaker.

## Custos de conformidade

Embora EC2 e EKS sejam seguros e compatíveis, o Amazon SageMaker é compatível com HIPAA, PCI, SOC, GDPR, ISO e FEDRAMP para cargas de trabalho de ML. O Amazon SageMaker também é compatível com endpoints FIPS. Embora a maioria das empresas tenha uma equipe central de auditoria e conformidade que estabelece os requisitos e gerencia o processo de conformidade, há um trabalho adicional para o proprietário do serviço preparar artefatos de conformidade, criar ferramentas de nível de serviço para demonstrar essa conformidade e mantê-la continuamente. Com o Amazon SageMaker, isso é fornecido imediatamente, sem nenhum custo adicional. O Amazon SageMaker economiza 60 meses de esforço de engenharia para conformidade para empresas grandes e muito grandes, o que equivale a 627 mil USD em custos durante um período de três anos, em comparação com a criação desses recursos no EC2 ou no EKS. Para fins de modelagem, supõe-se que os requisitos de conformidade para pequenas e médias empresas sejam mais baixos, necessitando de aproximadamente 25% e 50% desses custos, respectivamente.

# Cenários considerados

Embora cada empresa seja única, esta postagem assume quatro cenários para determinar o TCO.

Cenário	Pequeno	Médio	Grande	Extragrande
Operando na Nuvem AWS	Sim	Sim	Sim	Sim
Número de cientistas de dados na empresa	5	15	50	250
Crescimento anual das cargas de trabalho de ML	15%	10%	5%	2,5%
Regiões em que a empresa opera	Vários	Vários	Vários	Vários
Combinação de cargas de trabalho: ML clássico (como Scikit-Learn ou XGoost)	95%	85%	50%	50%
Combinação de cargas de trabalho: aprendizagem profunda (como PyTorch ou TensorFlow)	5%	15%	50%	50%
Duração do trabalho de treinamento por cientista de dados	0,1	0,5	2,5	5
Duração total dos trabalhos de treinamento por dia útil (horas)	0,5	7,5	250	1.250
Número total de modelos implantados na produção	2	10	50	150
Armazenamento do S3 (GB/mês)	25	250	5.000	25.000
Armazenamento do EBS (GB/mês)	15	150	1.000	5.000
Dados transferidos pela rede (GB/mês)	1	10	500	2.500
Requisitos de segurança e conformidade	Baixo (25%)	Médio (50%)	Alto (100%)	Alto (100%)
Proporção de engenheiros e cientistas de dados (autogerenciados)	1:10	1:20	1:30	1:40
Proporção de engenheiros e cientistas de dados (Amazon SageMaker)	Um décimo das opções autogerenciadas			
Salário anual de engenheiro em período integral	125 mil USD	125 mil USD	125 mil USD	125 mil USD

# Cenário 1: empresas de pequeno porte

O TCO do Amazon SageMaker para uma empresa de pequeno porte com cinco cientistas de dados é 96% menor que o EC2 e 96% menor que o EKS. As economias de TCO são de 97% no Ano 1, 94% no Ano 2 e 94% no Ano 3. A tabela a seguir apresenta um resumo da análise.

Análise de TCO de três anos					
Cenário pequeno	Amazon SageMaker	Opções "Faça você mesmo" no Amazon EC2	K8s gerenciadas na AWS	TCO do SM vs. EC2 "Faça você mesmo"	TCO do SM vs. K8s gerenciadas
<b>CUSTOS TOTAIS (Criação + Treino + Implantação)</b>	<b>54.831 USD</b>	<b>1.322.754 USD</b>	<b>1.284.710 USD</b>	<b>-96%</b>	<b>-96%</b>
<b>Custos totais da criação</b>	<b>16.124 USD</b>	<b>352.814 USD</b>	<b>352.814 USD</b>	<b>-95%</b>	<b>-95%</b>
<b>Custos totais de infraestrutura</b>	8.889 USD	6.356 USD	6.356 USD	40%	40%
<b>Custos operacionais</b>	7.234 USD	217.031 USD	217.031 USD	-97%	-97%
<b>Custos totais de segurança e conformidade</b>	- USD	129.427 USD	129.427 USD	-100%	-100%
<b>Custos totais de treinamento</b>	<b>14.681 USD</b>	<b>575.602 USD</b>	<b>573.446 USD</b>	<b>-97%</b>	<b>-97%</b>
<b>Custos totais de infraestrutura</b>	7.447 USD	229.144 USD	226.988 USD	-97%	-97%
<b>Custos operacionais</b>	7.234 USD	217.031 USD	217.031 USD	-97%	-97%
<b>Custos totais de segurança e conformidade</b>	- USD	129.427 USD	129.427 USD	-100%	-100%
<b>Custos totais de implantação</b>	<b>24.026 USD</b>	<b>394.337 USD</b>	<b>358.450 USD</b>	<b>-94%</b>	<b>-93%</b>
<b>Custos totais de infraestrutura</b>	16.792 USD	47.879 USD	11.991 USD	-65%	40%
<b>Custos operacionais</b>	7.234 USD	217.031 USD	217.031 USD	-97%	-97%
<b>Custos de segurança e conformidade</b>	- USD	129.427 USD	129.427 USD	-100%	-100%

## Os detalhes da criação são os seguintes:

- Há cinco cientistas de dados na equipe, cada um executando uma instância de bloco de anotações do T3.medium.
- O custo por instância por hora é de 0,0582 USD no Amazon SageMaker e 0,0416 USD no EC2 e EKS.



### **Os detalhes do treinamento são os seguintes:**

- 0,1 hora de treinamento por cientista de dados por dia útil. As instâncias de treinamento são uma combinação de M5.xlarge e P3.2xlarge, com base na combinação de ML clássico e cargas de trabalho de aprendizagem profunda para o cenário.
- O Amazon SageMaker não exige que você gerencie nenhum cluster e você paga apenas pela duração da execução dos trabalhos de treinamento. Para o EC2, cada cientista de dados executa uma instância dedicada. Com o EKS, você pode compartilhar uma instância entre quatro cientistas de dados, embora o cluster precise ficar continuamente em execução.
- O custo combinado por instância por hora é de 0,87 USD no Amazon SageMaker e 0,62 USD no EC2 e EKS.

### **Estes são os detalhes da implantação:**

- Dois modelos de ML implantados na produção. Cada modelo implantado é executado em duas instâncias. As instâncias são uma combinação de M5.xlarge e P3.2xlarge, com base na combinação de ML clássico e cargas de trabalho de aprendizagem profunda.
- Você pode implantar quatro modelos em uma instância com endpoints de vários modelos do Amazon SageMaker e quatro modelos em uma instância com EKS. Cada modelo implantado no EC2 exige instâncias dedicadas.
- O custo combinado por instância por hora é de 0,87 USD no Amazon SageMaker e 0,62 USD no EC2 e EKS.

### **Os detalhes a seguir se aplicam a todas as fases:**

- Os custos de segurança e conformidade são aproximadamente 25% dos custos incorridos por grandes empresas. Esse número é menor porque as pequenas empresas têm menos requisitos de segurança e conformidade.
- Os custos de armazenamento do S3 são de 0,023 USD por GB/mês.
- Os custos por volume do EBS são de 0,14 USD no Amazon SageMaker e 0,10 USD no EC2 e no EKS.
- Os custos de rede são de 0,016 USD no Amazon SageMaker e de 0,001 USD no EC2 e no EKS.

## Cenário 2: empresas de médio porte

O TCO do Amazon SageMaker para uma empresa de médio porte com 15 cientistas de dados é 91% menor que o EC2 e 90% menor que o EKS. As economias de TCO são de 94% no Ano 1, 86% no Ano 2 e 85% no Ano 3. A tabela a seguir apresenta um resumo da análise.

Análise de TCO de três anos					
Cenário médio	Amazon SageMaker	Opções "Faça você mesmo" no Amazon EC2	K8s gerenciadas na AWS	TCO do SM vs. EC2 "Faça você mesmo"	TCO do SM vs. K8s gerenciadas
<b>CUSTOS TOTAIS (Criação + Treino + Implantação)</b>	<b>213.233 USD</b>	<b>2.501.227 USD</b>	<b>2.189.631 USD</b>	<b>-91%</b>	<b>-90%</b>
<b>Custos totais da criação</b>	<b>36.013 USD</b>	<b>587.536 USD</b>	<b>587.536 USD</b>	<b>-94%</b>	<b>-94%</b>
<b>Custos totais de infraestrutura</b>	25.669 USD	18.369 USD	18.369 USD	40%	40%
<b>Custos operacionais</b>	10.344 USD	310.313 USD	310.313 USD	-97%	-97%
<b>Custos totais de segurança e conformidade</b>	- USD	258.854 USD	258.854 USD	-100%	-100%
<b>Custos totais de treinamento</b>	<b>26.313 USD</b>	<b>943.737 USD</b>	<b>932.525 USD</b>	<b>-97%</b>	<b>-97%</b>
<b>Custos totais de infraestrutura</b>	15.969 USD	374.570 USD	363.359 USD	-96%	-96%
<b>Custos operacionais</b>	10.344 USD	310.313 USD	310.313 USD	-97%	-97%
<b>Custos totais de segurança e conformidade</b>	- USD	258.854 USD	258.854 USD	-100%	-100%
<b>Custos totais de implantação</b>	<b>150.906 USD</b>	<b>969.954 USD</b>	<b>669.570 USD</b>	<b>-84%</b>	<b>-77%</b>
<b>Custos totais de infraestrutura</b>	140.563 USD	400.788 USD	100.404 USD	-65%	40%
<b>Custos operacionais</b>	10.344 USD	310.313 USD	310.313 USD	-97%	-97%
<b>Custos de segurança e conformidade</b>	- USD	258.854 USD	258.854 USD	-100%	-100%

### Os detalhes da criação são os seguintes:

- Há 15 cientistas de dados na equipe, cada um executando uma instância de bloco de anotações do T3.medium.
- O custo por instância por hora é de 0,0582 USD no Amazon SageMaker e 0,0416 USD no EC2 e EKS.



### **Os detalhes do treinamento são os seguintes:**

- 0,5 hora de treinamento por cientista de dados por dia útil. As instâncias de treinamento são uma combinação de M5.xlarge e P3.2xlarge, com base na combinação de ML clássico e cargas de trabalho de aprendizagem profunda para o cenário.
- O Amazon SageMaker não exige que você gerencie nenhum cluster e você paga apenas pela duração da execução dos trabalhos de treinamento. Para o EC2, cada cientista de dados executa uma instância dedicada. Com o EKS, você pode compartilhar uma instância entre quatro cientistas de dados, embora o cluster precise ficar continuamente em execução.
- O custo combinado por instância por hora é de 0,87 USD no Amazon SageMaker e 0,62 USD no EC2 e EKS.

### **Estes são os detalhes da implantação:**

- 150 modelos de ML implantados na produção. Cada modelo implantado é executado em duas instâncias. As instâncias são uma combinação de M5.xlarge e P3.2xlarge, com base na combinação de ML clássico e cargas de trabalho de aprendizagem profunda.
- Você pode implantar quatro modelos em uma instância com endpoints de vários modelos do Amazon SageMaker e quatro modelos em uma instância com EKS. Cada modelo implantado no EC2 exige instâncias dedicadas.
- O custo combinado por instância por hora é de 0,87 USD no Amazon SageMaker e 0,62 USD no EC2 e EKS.

### **Os detalhes a seguir se aplicam a todas as fases:**

- Os custos de segurança e conformidade são aproximadamente 50% dos custos assumidos por grandes empresas. Esse número é menor porque as empresas de médio porte têm poucos requisitos de segurança e conformidade.
- Os custos de armazenamento do S3 são de 0,023 USD por GB/mês.
- Os custos por volume do EBS são de 0,14 USD no Amazon SageMaker e 0,10 USD no EC2 e no EKS.
- Os custos de rede são de 0,016 USD no Amazon SageMaker e de 0,001 USD no EC2 e no EKS.

## Cenário 3: empresas de grande porte

O TCO do Amazon SageMaker para uma empresa de grande porte com 50 cientistas de dados é 78% menor que o EC2 e 65% menor que o EKS. As economias de TCO são de 73% no Ano 1, 62% no Ano 2 e 51% no Ano 3. A tabela a seguir apresenta um resumo da análise.

Análise de TCO de três anos					
Cenário grande	Amazon SageMaker	Opções "Faça você mesmo" no Amazon EC2	K8s gerenciadas na AWS	TCO do SM vs. EC2 "Faça você mesmo"	TCO do SM vs. K8s gerenciadas
<b>CUSTOS TOTAIS (Criação + Treino + Implantação)</b>	<b>2.062.111 USD</b>	<b>9.502.477 USD</b>	<b>5.866.850 USD</b>	<b>-78%</b>	<b>-65%</b>
<b>Custos totais da criação</b>	<b>105.571 USD</b>	<b>1.234.703 USD</b>	<b>1.234.683 USD</b>	<b>-91%</b>	<b>-91%</b>
<b>Custos totais de infraestrutura</b>	83.679 USD	60.224 USD	60.204 USD	39%	39%
<b>Custos operacionais</b>	21.892 USD	656.771 USD	656.771 USD	-97%	-97%
<b>Custos totais de segurança e conformidade</b>	- USD	517.708 USD	517.708 USD	-100%	-100%
<b>Custos totais de treinamento</b>	<b>277.127 USD</b>	<b>2.364.422 USD</b>	<b>2.273.403 USD</b>	<b>-88%</b>	<b>-88%</b>
<b>Custos totais de infraestrutura</b>	255.235 USD	1.189.943 USD	1.098.924 USD	-79%	-77%
<b>Custos operacionais</b>	21.892 USD	656.771 USD	656.771 USD	-97%	-97%
<b>Custos totais de segurança e conformidade</b>	- USD	517.708 USD	517.708 USD	-100%	-100%
<b>Custos totais de implantação</b>	<b>1.679.412 USD</b>	<b>5.903.351 USD</b>	<b>2.358.764 USD</b>	<b>-72%</b>	<b>-29%</b>
<b>Custos totais de infraestrutura</b>	1.657.520 USD	4.728.872 USD	1.184.285 USD	-65%	40%
<b>Custos operacionais</b>	21.892 USD	656.771 USD	656.771 USD	-97%	-97%
<b>Custos de segurança e conformidade</b>	- USD	517.708 USD	517.708 USD	-100%	-100%

### Os detalhes da criação são os seguintes:

- Há 50 cientistas de dados na equipe, cada um executando uma instância de bloco de anotações do T3.medium.
- O custo por instância por hora é de 0,0582 USD no Amazon SageMaker e 0,0416 USD no EC2 e EKS.

### **Os detalhes do treinamento são os seguintes:**

- 2,5 horas de treinamento por cientista de dados por dia útil. As instâncias de treinamento são uma combinação de M5.xlarge e P3.2xlarge, com base na combinação de ML clássico e cargas de trabalho de aprendizagem profunda para o cenário.
- O Amazon SageMaker não exige que você gerencie nenhum cluster e você paga apenas pela duração da execução dos trabalhos de treinamento. Para o EC2, cada cientista de dados executa uma instância dedicada. Com o EKS, você pode compartilhar uma instância entre quatro cientistas de dados, embora o cluster precise ficar continuamente em execução.
- O custo combinado por instância por hora é de 0,87 USD no Amazon SageMaker e 0,62 USD no EC2 e EKS.

### **Estes são os detalhes da implantação:**

- 50 modelos de ML implantados na produção. Cada modelo implantado é executado em duas instâncias. As instâncias são uma combinação de M5.xlarge e P3.2xlarge, com base na combinação de ML clássico e cargas de trabalho de aprendizagem profunda.
- Você pode implantar quatro modelos em uma instância com endpoints de vários modelos do Amazon SageMaker e quatro modelos em uma instância com EKS. Cada modelo implantado no EC2 exige instâncias dedicadas.
- O custo combinado por instância por hora é de 0,87 USD no Amazon SageMaker e 0,62 USD no EC2 e EKS.

### **Os detalhes a seguir se aplicam a todas as fases:**

- Os custos de segurança e conformidade incorridos somam 100% dos custos.
- Os custos de armazenamento do S3 são de 0,023 USD por GB/mês.
- Os custos por volume do EBS são de 0,14 USD no Amazon SageMaker e 0,10 USD no EC2 e no EKS.
- Os custos de rede são de 0,016 USD no Amazon SageMaker e de 0,001 USD no EC2 e no EKS.

## Cenário 4: empresas de porte muito grande

O TCO do Amazon SageMaker para uma empresa de porte muito grande com 250 cientistas de dados é 72% menor que o EC2 e 54% menor que o EKS. As economias de TCO são de 55% no Ano 1, 43% no Ano 2 e 43% no Ano 3. A tabela a seguir apresenta um resumo da análise.

Análise de TCO de três anos					
Cenário muito grande	Amazon SageMaker	Opções "Faça você mesmo" no Amazon EC2	K8s gerenciadas no AWS	TCO do SM vs. EC2 "Faça você mesmo"	TCO do SM vs. K8s gerenciadas
<b>CUSTOS TOTAIS (Criação + Treino + Implantação)</b>	<b>7.738.746 USD</b>	<b>27.551.979 USD</b>	<b>17.003.445 USD</b>	<b>-72%</b>	<b>-54%</b>
<b>Custos totais da criação</b>	<b>488.285 USD</b>	<b>3.214.319 USD</b>	<b>3.214.218 USD</b>	<b>-85%</b>	<b>-85%</b>
<b>Custos totais de infraestrutura</b>	408.191 USD	293.778 USD	293.678 USD	39%	39%
<b>Custos operacionais</b>	80.094 USD	2.402.832 USD	2.402.832 USD	-97%	-97%
<b>Custos totais de segurança e conformidade</b>	- USD	517.708 USD	517.708 USD	-100%	-100%
<b>Custos totais de treinamento</b>	<b>2.436.221 USD</b>	<b>7.924.226 USD</b>	<b>7.485.380 USD</b>	<b>-69%</b>	<b>-67%</b>
<b>Custos totais de infraestrutura</b>	2.356.126 USD	5.003.686 USD	4.564.840 USD	-53%	-48%
<b>Custos operacionais</b>	80.094 USD	2.402.832 USD	2.402.832 USD	-97%	-97%
<b>Custos totais de segurança e conformidade</b>	- USD	517.708 USD	517.708 USD	-100%	-100%
<b>Custos totais de implantação</b>	<b>4.814.240 USD</b>	<b>16.413.434 USD</b>	<b>6.303.846 USD</b>	<b>-71%</b>	<b>-24%</b>
<b>Custos totais de infraestrutura</b>	4.734.146 USD	13.492.894 USD	3.383.305 USD	-65%	40%
<b>Custos operacionais</b>	80.094 USD	2.402.832 USD	2.402.832 USD	-97%	-97%
<b>Custos de segurança e conformidade</b>	- USD	517.708 USD	517.708 USD	-100%	-100%

### Os detalhes da criação são os seguintes:

- Há 250 cientistas de dados na equipe, cada um executando uma instância de bloco de anotações do T3.medium.
- O custo por instância por hora é de 0,0582 USD no Amazon SageMaker e 0,0416 USD no EC2 e EKS.

### **Os detalhes do treinamento são os seguintes:**

- 5 horas de treinamento por cientista de dados por dia útil. As instâncias de treinamento são uma combinação de M5.xlarge e P3.2xlarge, com base na combinação de ML clássico e cargas de trabalho de aprendizagem profunda para o cenário.
- O Amazon SageMaker não exige que você gerencie nenhum cluster e você paga apenas pela duração da execução dos trabalhos de treinamento. Para o EC2, cada cientista de dados executa uma instância dedicada. Com o EKS, você pode compartilhar uma instância entre quatro cientistas de dados, embora o cluster precise ficar continuamente em execução.
- O custo combinado por instância por hora é de 0,87 USD no Amazon SageMaker e 0,62 USD no EC2 e EKS.

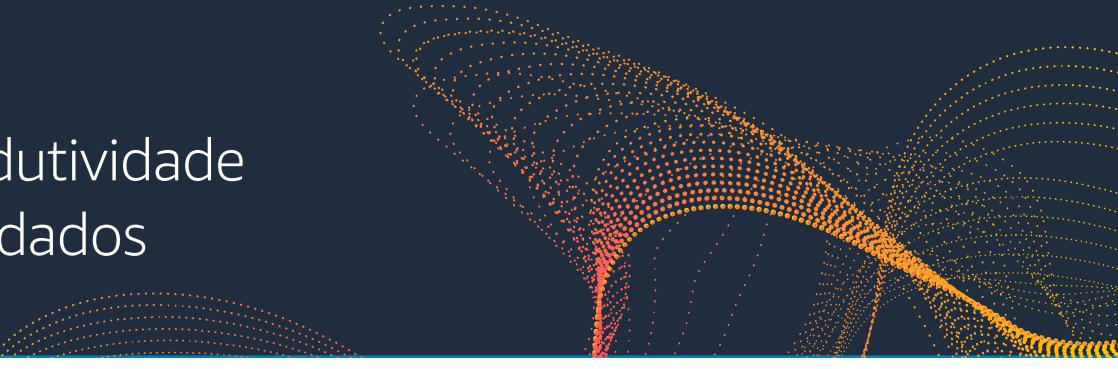
### **Estes são os detalhes da implantação:**

- 150 modelos de ML implantados na produção. Cada modelo implantado é executado em duas instâncias. As instâncias são uma combinação de M5.xlarge e P3.2xlarge, com base na combinação de ML clássico e cargas de trabalho de aprendizagem profunda.
- Quatro modelos podem ser implantados em uma instância com os endpoints de vários modelos do Amazon SageMaker. Quatro modelos podem ser implantados em uma instância com o EKS. Cada modelo implantado no EC2 exige instâncias dedicadas.
- O custo combinado por instância por hora é de 0,87 USD no Amazon SageMaker e 0,62 USD no EC2 e EKS.

### **Os detalhes a seguir se aplicam a todas as fases:**

- Os custos de segurança e conformidade incorridos somam 100% dos custos.
- Os custos de armazenamento do S3 são de 0,023 USD por GB/mês.
- Os custos por volume do EBS são de 0,14 USD no Amazon SageMaker e 0,10 USD no EC2 e no EKS.
- Os custos de rede são de 0,016 USD no Amazon SageMaker e de 0,001 USD no EC2 e no EKS.

# Ganhos de produtividade do cientista de dados



Atualmente, cientistas de dados e desenvolvedores de ML não são fáceis de encontrar. Não é apenas caro contratar o candidato certo com conhecimentos de ML, mas também pode ser extremamente difícil retê-lo devido à alta demanda por essas funções. Além disso, a maturidade dos processos de infraestrutura e desenvolvimento de ML limita frequentemente a produtividade de um cientista de dados. Você gasta grande parte do seu tempo em tarefas manuais, como configurar infraestrutura e ambientes, tarefas repetitivas com baixa automação, iterando nas versões do modelo, gerenciando os ciclos de vida do modelo e reescrevendo o código para torná-lo pronto para a produção. O Amazon SageMaker fornece componentes pré-criados e recursos totalmente gerenciados que aumentam a produtividade do cientista de dados em até 10 vezes.

O Amazon SageMaker facilita a criação, o treinamento, o ajuste e a implantação de modelos em escala. Ele reúne todas as ferramentas para executar, depurar e automatizar as etapas completas de criação e gerenciamento de um modelo de ML. Você pode escrever código, iniciar e acompanhar experimentos de ML, modelos de depuração e monitorar a qualidade do modelo, tudo por meio de uma única interface unificada, melhorando assim a produtividade. Esta publicação assume as seguintes hipóteses sobre cada fase.

## Fase de construção

Durante a fase de construção, você pode explorar dados, experimentar várias estruturas de ML, algoritmos e criar modelos em pequenas amostras de dados. O Amazon SageMaker fornece blocos de anotações Jupyter pré-criados e totalmente gerenciados que precisam de configuração zero, são pré-configurados com as bibliotecas ML mais recentes e populares e são fáceis de compartilhar com os colegas. A preparação de dados é um desafio importante e um processo demorado para os cientistas de dados. O Amazon SageMaker fornece ferramentas visuais de preparação de dados, além da capacidade de executar tarefas de análise ad-hoc, ETL e Spark em uma infraestrutura totalmente gerenciada e distribuída. Além disso, para reduzir o tempo necessário para criar modelos de alta qualidade, o Amazon SageMaker fornece recursos de AutoML e 17 algoritmos de ML pré-criados que cobrem os casos de uso mais comuns de ML em aprendizagem supervisionada, não supervisionada e de reforço.

## Fase de treinamento

Durante a fase de treinamento, você pode experimentar técnicas de processamento de dados, vários algoritmos de ML e hiperparâmetros do modelo. O número de iterações e trabalhos de treinamento necessários antes da implantação bem-sucedida de um modelo pode variar de centenas a milhares. Os recursos de ajuste automático de modelo do Amazon SageMaker usam o ML para encontrar o melhor modelo com base em objetivos definidos pelo cliente e reduzir o tempo necessário para obter modelos de alta qualidade. Além disso, os recursos de depuração no Amazon SageMaker permitem detectar problemas nos trabalhos de treinamento antecipadamente, analisar melhor os modelos de aprendizagem profunda e obter mais informações sobre a performance do treinamento. Por fim, à medida que o número de projetos, testes e experimentos na organização aumenta para a ordem de milhares, os recursos de gerenciamento de experimentos do Amazon SageMaker e o IDE de ML facilitam a organização, o rastreamento e o gerenciamento de experimentos e trabalhos de treinamento em ML no tempo e nas organizações.

## Fase de implantação

Durante a fase de implantação, você pode colocar seus modelos de ML em produção e fazer inferências sobre dados invisíveis. O Amazon SageMaker permite a implantação do modelo com um clique, sem a necessidade de alterações no código. Você pode treinar os modelos no Amazon SageMaker ou em ambientes diferentes e criá-los usando modelos internos do Amazon SageMaker ou modelos personalizados. Você pode implantar os mesmos modelos para fazer inferências em baixa latência em tempo real ou para lotes de dados. Além disso, você pode implantar novos modelos perfeitamente, sem nenhum impacto na disponibilidade ou perda de performance. Por fim, o Amazon SageMaker salva automaticamente solicitações e respostas de inferência e analisa os dados coletados periodicamente para detectar desvios do modelo e problemas de qualidade de dados na produção. Esses monitoramento e alerta contínuos permitem executar ações preventivas, como treinar novamente os modelos sem criar ferramentas.

# Histórias de sucesso de clientes



A seguir, são apresentadas histórias de sucesso sobre como os clientes do Amazon SageMaker economizaram custos e aumentaram a produtividade.



A Coinbase usa modelos de ML no Amazon SageMaker para ajudar na prevenção de fraudes, verificação de identidade e conformidade em larga escala. **O uso do Amazon SageMaker reduziu o tempo de treinamento do modelo de 20 horas para 10 minutos.**



A Intuit desenvolveu modelos de ML que podem analisar um ano de transações bancárias para encontrar despesas comerciais dedutíveis para os clientes. **Usando o Amazon SageMaker, o Intuit reduziu o tempo de implantação do ML em 90%, de 6 meses para 1 semana.**



Usando o Amazon SageMaker, a NuData Security evita fraudes no cartão de crédito, analisando dados anônimos do usuário para detectar atividades anormais antes que ocorra uma transação fraudulenta. **Com o Amazon SageMaker, a NuData reduziu o tempo de desenvolvimento de ML em 60%, simplificou sua arquitetura de ML em 95% e trabalhou com um grande banco para bloquear passivamente quase 100% do tráfego de tentativas fraudulentas dentro da tolerância de atrito do consumidor do banco.**



Usando o Amazon SageMaker, a Voodoo pode decidir em tempo real qual anúncio será exibido aos seus jogadores e acionar seu endpoint mais de 100 milhões de vezes por mais de 30 milhões de usuários diariamente, representando quase um bilhão de previsões por dia. **Com o machine learning da AWS, a Voodoo colocou um modelo preciso em produção em menos de uma semana** com o suporte de uma equipe pequena, e continuou o desenvolvimento com base nele à medida que a equipe e os negócios aumentavam.



Usando o TensorFlow no Amazon SageMaker, a Siemens Financial Services desenvolveu um modelo de PNL para extrair informações críticas a fim de acelerar a due diligence de investimentos, **reduzindo o tempo para resumir os documentos de diligência de 12 horas para 30 segundos.**



A Celgene usa o Apache MXNet no Amazon SageMaker para previsão de toxicologia a fim de analisar virtualmente os impactos biológicos de possíveis medicamentos sem colocar os pacientes em risco. **Um modelo que anteriormente levava dois meses para treinar já pode ser treinado em quatro horas.**



A ADP usa o machine learning da AWS, incluindo o Amazon SageMaker, para identificar rapidamente os padrões da força de trabalho e prever os resultados antes que eles ocorram, como por exemplo a rotatividade de funcionários ou o impacto de um aumento na remuneração. **A ADP reduziu o tempo de implantação dos modelos de machine learning de duas semanas para apenas um dia.**

# Conclusão

O Amazon SageMaker é um serviço modular de ML de ponta a ponta, totalmente gerenciado, que permite criar, treinar, ajustar e implantar modelos em escala. Ele elimina os custos indiretos de provisionamento de hardware e software de sistemas, gerenciamento de infraestrutura e operações, e criação de segurança e conformidade para cargas de trabalho de ML. Isso permite que você se concentre nos problemas de negócios, desenvolva com novas técnicas e leve os projetos de ML da ideia à produção mais rapidamente. O custo total de propriedade do Amazon SageMaker em um período de três anos é 54% menor em comparação com outras opções de ML baseadas na nuvem, como as opções EC2 autogerenciadas e as opções gerenciadas de Kubernetes (K8s) na AWS, como o EKS. Além do TCO mais baixo, os recursos totalmente gerenciados e integrados do Amazon SageMaker permitem colocar as ideias de ML em produção mais rapidamente e melhorar a produtividade do cientista de dados em até 10 vezes.