

The total cost of ownership (TCO) of Amazon SageMaker

Last updated: February 2, 2020

Please send questions to:
sagemaker_inquiries@amazon.com

Table of Contents

- Amazon SageMaker Total Cost of Ownership (TCO)2**
- Services considered for the TCO analysis3**
- Scenarios considered for the TCO analysis4**
- Costs considered for the TCO analysis4**
- Operational costs5
- Security and compliance costs6
- Infrastructure costs.....7
- TCO analysis by scenario..... 11**
- Scenario 1: Small-sized business 11
- Scenario 2: Medium-sized business..... 17
- Scenario 3: Large-sized business 23
- Scenario 4: Extra-large-sized business 29
- Data scientist productivity gains and time savings..... 34**
- Success stories 36**
- Conclusion..... 38**

Amazon SageMaker Total Cost of Ownership (TCO)

Amazon SageMaker is a fully managed machine learning (ML) service to build, train, and deploy models at scale and at significantly lower costs compared to other ML options we offer. The total cost of ownership (TCO) of Amazon SageMaker over a three-year horizon can be over 54% lower compared to other self-managed cloud-based ML options such as Amazon Elastic Compute Cloud EC2 (Amazon EC2) and Amazon Elastic Kubernetes Service (Amazon EKS). In addition to the lower TCO, Amazon SageMaker’s fully managed and integrated features let you put ML ideas into production faster and improve data scientist productivity by up to 10 times.



Coinbase uses ML models on Amazon SageMaker to help with fraud prevention, identity verification, and large-scale compliance. **Using Amazon SageMaker, Coinbase reduced model training times from 20 hours to 10 minutes.**



Intuit developed ML models that can pull a year’s worth of bank transactions to find deductible business expenses for customers. **Using Amazon SageMaker, Intuit reduced ML deployment time by 90%, from six months to one week.**

Teams of all sizes benefit from significantly lower TCO when using Amazon SageMaker. For example, over three-years, a small team of 5 data scientists can realize up to 90% lower TCO using Amazon SageMaker versus building and maintaining their own ML services on EC2 or EKS. Medium-sized teams of 15 data scientists can realize up to 87% lower TCO with Amazon SageMaker compared to EC2 and up to 85% lower TCO compared to EKS. Large teams of 50 data scientists can realize up to 79% lower TCO compared to EC2 and up to 65% lower TCO compared to EKS. Even larger teams of 250 data scientists can realize up to 77% lower TCO with Amazon SageMaker compared to EC2, and up to 54% lower TCO compared to EKS on AWS.

Overall summary		Amazon SageMaker three-year TCO Savings	
		Compared to EC2	Compared to EKS
Small Scenario	5 Data Scientists	-90%	-90%
Medium Scenario	15 Data Scientists	-87%	-85%
Large Scenario	50 Data Scientists	-79%	-65%
Extra-Large Scenario	250 Data Scientists	-77%	-54%

Typically, the TCO for Amazon SageMaker is lower in the first year compared to EC2 or EKS options due to the higher investments required to build security and compliance features which come out of the box with Amazon SageMaker. The TCO for Amazon SageMaker continues to remain significantly lower than other solutions in subsequent years. This is due to the much greater infrastructure utilization Amazon SageMaker achieves out of the box compared to EC2 or EKS, and the higher ongoing costs of maintaining the core infrastructure, security, and compliance for the self-managed EC2 or managed EKS options.

In the TCO analysis, we evaluated infrastructure costs (compute, storage, and network), operational costs, and security and compliance costs for each step of the ML workflow across different team sizes (small, medium, large, extra-large.) During ML model building, you incur costs to explore and preprocess data and experiment with ML frameworks and algorithms. During training, you incur costs for training ML models as well as tuning ML model hyperparameters. Finally, during ML model deployment, you incur costs as your model makes inferences on unseen data. Across each step of the workflow, we also factor in the costs to employ engineers and maintain security and compliance.

Services considered for the TCO analysis

The TCO analysis has considered the following options:

1. **Self-managed on EC2** - This option gives you a DIY approach to ML using AWS services such as AWS Batch and Amazon ECS. You take on the responsibility of provisioning and managing EC2 instances, including instance failure recovery, patching, and automatic scaling. You use the DLAMI with the ML frameworks and libraries pre-built but need to optimize data access to get high throughput, optimize your setup for scale, and enable distributed training. In addition, you need to build and maintain the required security and compliance features for your ML workloads.
2. **Managed Kubernetes on AWS** - Services such as Amazon EKS make it easy to deploy, manage, and scale containerized workloads on EC2. You need to manage your own cluster and tune its performance and utilization based on the memory, compute, and network requirements for your workloads. In addition, you need to build the right level of security, compliance, and availability for your ML workloads. While there are open-source tools such as Kubeflow that make running ML workloads easier, you still incur infrastructure management costs because of the expertise required to build, manage, and secure a Kubernetes cluster is much higher and less readily available.
3. **Fully managed ML service** – Amazon SageMaker is fully managed so you don't need to build, manage, or maintain any infrastructure, and tooling to build, train, or deploy ML models. You don't have to manage any instances, patching, upgrades, or build automatic scaling and load balancing capabilities to scale your workloads. This reduces your required operational support to less than one-tenth compared to self-managed solutions. You also get high instance utilization and only pay for compute based on your usage. Amazon SageMaker has built-in security and compliance for ML workloads and does not require you to invest in additional security for running your workloads.

Scenarios considered for the TCO analysis

While every business is unique, we considered four business scenarios to determine the TCO.

Scenario	Small	Medium	Large	X-Large
Operating in the AWS Cloud	Yes	Yes	Yes	Yes
Regions the company operates in	Multiple	Multiple	Multiple	Multiple
Number of data scientists in the company	5	15	50	250
Yearly growth of ML workloads and infrastructure usage	15%	10%	5%	2.5%
Mix of Classic ML workloads (Scikit-Learn, XGBoost etc.)	95%	85%	50%	50%
Mix of Deep learning workloads (PyTorch, TensorFlow etc.)	5%	15%	50%	50%
Total duration of training jobs per business day (hours)	1	5	20	110
Total number of models deployed in production	2	5	25	100
Amazon Simple Storage Service (GB/month)	75	1,500	15,000	75,000
Amazon Elastic Block Store (GB/month)	35	150	1,500	7,500
Data transferred over network (GB/month)	100	500	7,500	25,000
Security & compliance features (% of SageMaker features)	15%	30%	60%	100%
Ratio of engineers to data scientists (self-managed)	1:10	1:20	1:30	1:40
Ratio of engineers to data scientists (Amazon SageMaker)	1:100	1:200	1:300	1:400
Annual salary for full-time software engineer	\$100K	\$100K	\$100K	\$100K

Costs considered for the TCO analysis

Calculating the TCO for Amazon SageMaker included evaluating the cost of running infrastructure, managing operations, and building the security and compliance for running ML workloads. We evaluated these costs across each of the three ML phases over a three-year period:

- **Build:** Explore, preprocess data, experiment with ML frameworks and algorithms using notebooks.
- **Train:** Train ML models and tune model hyperparameters on large datasets.
- **Deploy:** Put ML models into production and make inferences on unseen data.

We know that the costs for ML can vary depending on the type of model you use. Our TCO analysis isn't based on one particular ML framework, algorithm, or model. Instead, it takes a common mix of both machine learning and deep learning models that we have seen in production across our customers. We also assume yearly growth in ML workloads and team sizes for each scenario. We would also like to note that some line items may not line up exactly because of rounding differences.

Operational costs

The operational costs are the combined costs of operating, monitoring, and maintaining the machine learning infrastructure. The operational costs are calculated as the number of engineers required based on each scenario and the annual salary of engineers, aggregated over a three-year period. Customers using self-managed ML solutions on EC2 and EKS need to build operational tooling themselves. Customers using Amazon SageMaker incur significantly less operational costs because Amazon SageMaker is fully managed and provides these capabilities out-of-the-box. Customers do not need to provision instances, monitor instance health, manage security updates or patches, emit operational metrics, and do not need to build monitoring for their ML workloads.

A report from ZipRecruiter¹ states that the average annual pay for a software engineer in the United States is \$98,846 as of December 29th, 2019. The report also highlights \$79,000 as the 25th percentile of the salary range and \$115,500 as the 75th percentile of the salary range. Based on over 26,500 salaries submitted, Indeed reports² \$107,698 as the average annual base salary for software engineers in the United States with an additional \$4K in cash bonus. For the TCO analysis, we've assumed that the average annual software engineer salary is \$100K for a three-year period.

For the small scenario, we assume there are 5 data scientists on the team and 1 engineer is required to support a team of 10 data scientists for EC2 and EKS. With the annual salary assumption of \$100K and 15% yearly growth in the number of data scientists, the total operations costs per year are \$50,000 in year 1, \$57,500 in year 2, and \$66,125 in year 3, totaling \$173,625 over a three-year period. For Amazon SageMaker, we assume the engineering support required is one-tenth of the support needed for the other DIY options due to the fully managed nature of the service. The total operational costs for Amazon SageMaker are \$5,000 in year 1, \$5,750 in year 2, and \$6,613 in year 3, totaling \$17,363 over a three-year period. These costs are equally divided and allocated to each of the three ML phases.

For the medium scenario, we assume there are 15 data scientists on the team and 1 engineer is required to support a team of 20 data scientists for EC2 and EKS. With the annual salary assumption of \$100K and 10% yearly growth in the number of data scientists, the total operations costs per year are \$75,000 in year 1, \$82,500 in year 2, and \$90,750 in year 3, totaling \$248,250 over a three-year period. For Amazon SageMaker, we assume the engineering support required is one-tenth of the support needed for the other DIY options due to the fully managed nature of the service. The total operational costs for Amazon SageMaker are \$7,500 in year 1, \$8,250 in year 2, and \$9,075 in year 3, totaling \$24,825 over a three-year period. These costs are equally divided and allocated to each of the three ML phases.

For the large scenario, we assume there are 50 data scientists on the team and 1 operations engineer is required to support a team of 30 data scientists for EC2 and EKS. With the annual

¹ <https://www.ziprecruiter.com/Salaries/Software-Engineer-Salary>

² <https://www.indeed.com/career/software-engineer/salaries>

salary assumption of \$100K and 5% yearly growth in the number of data scientists, the total operations costs per year are \$166,667 in year 1, \$175,000 in year 2, and \$183,750 in year 3, totaling \$525,417 over a three-year period. For Amazon SageMaker, we assume the engineering support required is one-tenth of the support needed for the other DIY options due to the fully managed nature of the service. The total operational costs for Amazon SageMaker are \$16,667 in year 1, \$17,500 in year 2, and \$18,375 in year 3, totaling \$52,542 over a three-year period. These costs are equally divided and allocated to each of the three ML phases.

For the extra-large scenario, we assume there are 250 data scientists on the team and 1 operations engineer is required to support a team of 40 data scientists for EC2 and EKS. With the annual salary assumption of \$100K and 2.5% yearly growth in the number of data scientists, the total operations costs per year are \$625,000 in year 1, \$640,625 in year 2, and \$656,641 in year 3, totaling \$1,922,266 over a three-year period. For Amazon SageMaker, we assume the operational engineering support required is one-tenth of the support needed for the other DIY options due to the fully managed nature of the service. The total operational costs for Amazon SageMaker are \$62,500 in year 1, \$64,063 in year 2, and \$65,664 in year 3, totaling \$192,227 over a three-year period. These costs are equally divided and allocated to each of the three ML phases.³

Security and compliance costs

ML workloads rely on large volumes of data, much of it sensitive or proprietary, to build and train models. When the models are in production, the ML models need to secure incoming inference requests as well as responses. Though EC2 and EKS are secure and compliant services, you need to build the security and compliance features required for ML, which incurs additional engineering costs. For example, you need to secure data at rest and in transit, control access, and maintain compliance when you use notebooks, train across multiple nodes, and run inferences on customer data. In contrast, Amazon SageMaker is designed to run ML applications at scale with the required security, compliance, and world-wide availability so you don't need to build these capabilities.

Security costs

Amazon SageMaker supports security with end-to-end encryption at rest and in transit, including encryption of the root volume and Amazon Elastic Block Store volume, Amazon Virtual Private Cloud support, AWS PrivateLink, customer-managed keys, AWS Identity and Access Management fine-grained access control, AWS CloudTrail audits, internode encryption for training, tag-based access control, network isolation, and Interactive Application Proxy.

All of these security features provided out of the box in Amazon SageMaker save businesses 91 development months of engineering effort over a three-year period. With the software engineer annual salary assumption of \$100K, this is equivalent to \$758,333 in cost savings when using Amazon SageMaker as compared to building these security features on top of EC2 and EKS.

³ Some line items may not completely match detailed breakouts due to rounding.

Compliance costs

Most companies have a central audit and compliance team that lays down the requirements for compliance and manages the compliance process. However, there is additional work for the service owner to prepare compliance artifacts, build service level tooling to demonstrate compliance, and maintain the compliance on an ongoing basis. With Amazon SageMaker, this is provided out of the box at no additional cost. Amazon SageMaker is a HIPAA eligible service, and is certified under PCI, SOC, GDPR, and ISO. Amazon SageMaker also supports FIPS endpoints. While EC2 and EKS are certified under numerous certification programs, you need to build this tooling yourself to achieve your desired compliance goals. All of the security features provided out of the box in Amazon SageMaker save businesses 61 development months of engineering effort over a three-year period. With the software engineer annual salary assumption of \$100K, this is equivalent to \$508,333 in cost savings when using Amazon SageMaker as compared to building compliance features in EC2 and EKS.

For modeling purposes, we assume that businesses would incur majority of these costs in year 1 to build the security and compliance features when using EC2 and EKS, and they incur a smaller portion of the costs in the next two years to maintain and improve these features. The total security and compliance cost savings with Amazon SageMaker over a three-year period are \$1,266,667 (\$758,333+\$508,333). For the TCO analysis, we assume that businesses have varying security and compliance requirements. We assume businesses in the small and medium scenarios needing to build approximately 15% and 30% of the security and compliance features provided out of the box by Amazon SageMaker respectively. We assume businesses in the large and extra-large scenarios needing to build 60% and 100% of the security and compliance features provided out of the box by Amazon SageMaker respectively.

Based on these assumptions, businesses in the small scenario incur \$190,000 in costs over a three-year period to build the required security and compliance features for their ML workloads. Businesses in the medium scenario incur \$380,000 in costs over a three-year period. Businesses in the large scenario incur \$760,000 in costs over a three-year period. Businesses in the extra-large scenario incur \$1,266,667 in costs over a three-year period. In contrast, since these features are provided out of the box in Amazon SageMaker, customers do not incur these costs⁴. These costs are equally divided and allocated to each of the three ML phases.

Infrastructure costs

The infrastructure costs are the combined costs for storage, network, and compute. We assume yearly growth in ML workloads, usage, and team sizes but not in software engineer salaries. The compute costs are calculated using on-demand instances and based on the usage assumptions for each scenario over a three-year period. The storage costs for Amazon Simple Storage Service (Amazon S3) and Amazon Elastic Block Store (Amazon EBS) are based on the monthly costs per GB and the usage assumptions for each scenario. The network costs

⁴ Client or application level security and compliance costs are not included in the TCO analysis

are the aggregated costs of data transferred in and out of the network based on the monthly costs per GB and the usage assumptions for each scenario.

Storage costs

We assume that customers' data for training ML models is stored in Amazon S3 (S3). The S3 pricing used for the TCO analysis is \$0.023 per GB⁵ and the S3 storage costs are the same across Amazon SageMaker, EC2, and EKS in all the four scenarios. In addition, customers use EBS General Purpose SSD (gp2) Volumes in each of the ML phases and pay based on their usage and per unit prices⁶. The EBS pricing is \$0.10 per GB-month of provisioned storage when customers use EC2 and EKS, and \$0.14 per GB-month of provisioned storage when customers use Amazon SageMaker. The storage costs for the TCO analysis include the S3 and EBS costs over a three-year period.

For the small scenario, we assume 75GB of S3 storage and 35GB of EBS storage per month, and a yearly growth rate of 15%. Over a three-year period the total storage cost for EC2 and EKS is \$218 and the cost for Amazon SageMaker is \$276. These costs are allocated equally to all three ML phases.

For the medium scenario, we assume 1.5TB of S3 storage and 150GB of EBS storage per month, and a yearly growth rate of 10%. Over a three-year period the total storage cost for EC2 and EKS is \$1,966 and the cost for Amazon SageMaker is \$2,204. These costs are allocated equally to all three ML phases.

For the large scenario, we assume 15TB of S3 storage and 1.5TB of EBS storage per month, and a yearly growth rate of 5%. Over a three-year period the total storage cost for EC2 and EKS is \$18,726 and the cost for Amazon SageMaker is \$20,996. These costs are allocated equally to all three ML phases.

For the extra-large scenario, we assume 75TB of S3 storage and 7.5TB of EBS storage per month, and a yearly growth rate of 2.5%. Over a three-year period the storage cost for EC2 and EKS is \$91,346 and the cost for Amazon SageMaker is \$102,418. These costs are allocated equally to all three ML phases.

Network costs

The network costs are applicable only in the deploy phase for data transferred in and out of the network over a three-year period. The network cost for EC2 and EKS is \$0.01 per GB and the network cost for Amazon SageMaker is \$0.016 per GB.

For the small scenario, we assume 100GB data transferred over the network in a month and a yearly growth rate of 15%. Over a three-year period total network cost for EC2 and EKS is \$48 and the cost for Amazon SageMaker is \$67.

⁵ S3 pricing is assumed steady at \$0.023 per GB for all four scenarios: Small, Medium, Large, and Extra-large.

⁶ We assume the usage is for the full month

For the medium scenario, we assume 500GB data transferred over the network in a month and a yearly growth rate of 10%. Over a three-year period total network cost for EC2 and EKS is \$227 and the cost for Amazon SageMaker is \$318.

For the large scenario, we assume 7.5TB data transferred over the network in a month and a yearly growth rate of 5%. Over a three-year period total network cost for EC2 and EKS is \$3,243 and the cost for Amazon SageMaker is \$4,540.

For the extra-large scenario, we assume 25TB data transferred over the network in a month and a yearly growth rate of 2.5%. Over a three-year period total network cost for EC2 and EKS is \$10,545 and the cost for Amazon SageMaker is \$14,763.

Compute costs

The compute costs are calculated differently for each of the three ML phases: Build, Train, and Deploy. The compute costs are calculated using on-demand instances and the usage assumptions for each scenario over a three-year period.

For the build phase, we assume that each data scientist runs a Notebook instance using the T3.medium instance type and runs it 24 hours a day for a year. The compute cost is calculated as the total notebook instances in use * price per hour * hours in a year aggregated over a three-year period. Even though customers can automate shutting down notebook instances in all three services, for the TCO analysis, we assume that these notebooks are running 24 hours a day.

For the train phase, customers using EC2 launch dedicated instances for each data scientist and the instances run during business hours⁷. Customers using EKS have higher instance utilization and can share an instance among four data scientists⁸, thereby requiring fewer instances to run during business hours. Compute costs for EC2 and EKS⁹ are calculated as the total training instances in use * cost per hour * hours in a year aggregated over a three-year period.

Amazon SageMaker Training sets up the infrastructure on your behalf when you start the training job and releases the instances once the training is complete. The compute costs are incurred only for the duration for which the training runs and aggregated across all training jobs over a three-year period. We assume that customers use M5.xlarge for training classic ML workloads and P3.2xlarge for training deep learning models. The instance cost per hour is a blended cost per hour based on the mix of classic ML and deep learning for each scenario and the price per hour of the respective instances.

For the deploy phase, we assume that models are deployed for real-time inferences and the underlying instances are available 24 hours a day and 7 days a week to serve inferences with low latency. Assuming each model requires one instance, EC2 customers require as many

⁷ We assume training is running during business hours in all scenarios, i.e. 8 hours each day for 5 days in a week.

⁸ Rounded up to whole numbers

⁹ We have not included the cost of running the cluster in EKS for the TCO analysis

instances as there are models deployed. EKS offers higher instance utilization and we assume that four models can be deployed on a single instance¹⁰.

Amazon SageMaker Multi-model endpoints enable customers to deploy multiple models on an endpoint deployed on one or more instances. Even though Amazon SageMaker Multi-model endpoints can scale to serve hundreds and thousands of models from the same endpoint, for the TCO calculation, we have assumed that four models can run on an endpoint deployed on a single instance. The compute costs are calculated as the number of instances in service * cost per hour * hours in a year aggregated over a three-year period. We assume that customers use M5.xlarge for deploying classic ML models and P3.2xlarge for deploying deep learning models. The instance cost per hour is a blended cost per hour based on the mix of classic ML and deep learning for each scenario and the price per hour of the respective instances.

¹⁰ Rounded up to whole number for EKS and Amazon SageMaker

TCO analysis by scenario

The following section presents the TCO for each of the 4 scenarios.

Scenario 1: Small-sized business

The TCO of Amazon SageMaker for a small-sized business with five data scientists is up to 90% lower than EC2 and up to 90% lower than EKS. The following table presents a summary of the analysis.

3 year TCO Analysis					
Small Scenario	Amazon SageMaker	Amazon EC2	Amazon EKS	Amazon SageMaker TCO v/s EC2	Amazon SageMaker TCO v/s EKS
TOTAL COSTS (Build + Train + Deploy)	\$39,658	\$409,985	\$383,628	-90%	-90%
Total Build Costs	\$15,056	\$127,840	\$127,840	-88%	-88%
Infrastructure Costs	\$9,269	\$6,632	\$6,632		
Operational Costs	\$5,788	\$57,875	\$57,875		
Security & Compliance Costs	\$ -	\$63,333	\$63,333		
Total Train Costs	\$6,304	\$134,011	\$125,524	-95%	-95%
Infrastructure Costs	\$516	\$12,802	\$4,316		
Operational Costs	\$5,788	\$57,875	\$57,875		
Security & Compliance Costs	\$ -	\$63,333	\$63,333		
Total Deploy Costs	\$18,298	\$148,134	\$130,264	-88%	-86%
Infrastructure Costs	\$12,510	\$26,926	\$9,055		
Operational Costs	\$5,788	\$57,875	\$57,875		
Security & Compliance Costs	\$ -	\$63,333	\$63,333		

Build costs

The build costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$218 and the cost for Amazon SageMaker is \$276. The storage costs are allocated equally among the three phases of ML and the storage cost for the build

phase over a three-year period for EC2 and EKS is \$73. The storage cost for the build phase over a three-year period for Amazon SageMaker is \$92.

- **Network costs:** There are no network costs for the build phase.
- **Compute costs:** The compute cost is calculated as the total notebook instances running * hours in a year * cost per hour aggregated across a three-year period. With a 15% yearly growth rate in the team size, the total notebook instances running for this scenario are 5 in year 1, 6 in year 2, and 7 in year 3¹¹. For the build phase, each data scientist is running one notebook instance of T3.medium for 24 hours every day of the year. The cost per instance per hour is \$0.0416 in EC2 and EKS and \$0.0582 in Amazon SageMaker.

The compute cost for EC2 is \$1,822 in year 1, \$2,186 in year 2, and \$2,551 in year 3, totaling \$6,559 over a three-year period.

The compute cost for EKS is \$1,822 in year 1, \$2,186 in year 2, and \$2,551 in year 3, totaling \$6,559 over a three-year period.

The compute cost for Amazon SageMaker is \$2,549 in year 1, \$3,059 in year 2, and \$3,569 in year 3, totaling \$9,177 over a three-year period.

- The total infrastructure cost is calculated as the sum of storage, network, and compute costs. The infrastructure cost for EC2 in the build phase is $\$73 + \$6,559 = \$6,632$. The infrastructure cost for EKS in the build phase is $\$73 + \$6,559 = \$6,632$. The total infrastructure cost for Amazon SageMaker in the build phase is $\$92 + \$9,177 = \$9,269$.

For the small scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$173,625 across all ML phases. For Amazon SageMaker, we previously calculated that total operational cost is \$17,363 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the build phase for EC2 and EKS over a three-year period is \$57,875. The operational cost for the build phase for Amazon SageMaker over a three-year period is \$5,788.

For the small scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$190,000 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the build phase for EC2 and EKS over a three-year period is \$63,333. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the build phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total build cost for EC2 over a three-year period is $\$6,632 + \$57,875 + \$63,333 = \$127,840$. The total build cost for EKS over a three-year period is $\$6,632 + \$57,875 + \$63,333 = \$127,840$. The total build cost for Amazon SageMaker over a three-year period is $\$9,269 + \$5,788 + \$0 = \$15,056$. The TCO of Amazon SageMaker for a small-sized business with five data scientists for the build phase is 88% lower than EC2 and 88% lower than EKS.

¹¹ Instance count and number of data scientists are rounded up

Train

The training costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$218 and the cost for Amazon SageMaker is \$276. The storage costs are allocated equally among the three phases of ML. The storage cost for the train phase over a three-year period for EC2 and EKS is \$73. The storage cost for the train phase over a three-year period for Amazon SageMaker is \$92.
- **Network costs:** There are no network costs for the train phase.
- **Compute costs:** The compute cost is calculated as the total duration for which instances are in use for training in a year * cost per instance hour aggregated across a three-year period. Based on the assumptions for this scenario, 95% of the workload mix is classic ML and the remaining 5% workload mix is deep learning. We are using M5.xlarge for training classic ML workloads and P3.2xlarge for training deep learning workloads. The cost per instance hour for m5.xlarge and p3.2xlarge in EC2 and EKS is \$0.192 and \$3.06 respectively. Based on the workload mix, the blended instance cost per hour in EC2 and EKS is $(0.95 * \$0.192) + (0.05 * \$3.06) = \$0.34$.
The cost per instance hour for ml.m5.xlarge and ml.p3.2xlarge in Amazon SageMaker is \$0.269 and \$4.284 respectively. The blended instance cost per hour in Amazon SageMaker is $(0.95 * \$0.269) + (0.05 * \$4.284) = \$0.47$.

For the small scenario we assume that the data scientists on the team may be training during business hours, i.e. 5 days each week and 8 hours each day for 52 weeks in a year¹². Based on the growth rate, the number of data scientists increase from 5 in year 1, to 6 data scientists in year 2 and 7 data scientists in year 3. We assume that the total training duration aggregated across all data scientists is 1 hour per business day in Year 1. With a yearly growth of 15% the total training duration aggregated across all data scientists is 1.2 hours per business day in Year 2, and 1.3 business hours per business day in Year 3.

For training using EC2, we assume that each data scientist needs to provision an instance that is running during business hours. As a result, the number of instances running with EC2 are 5 in year 1, 6 in year 2, and 7 in year 3. With the blended cost of instance per hour of \$0.34, the training cost is \$3,536 in year 1, \$4,243 in year 2, and \$4,950 in year 3, totaling \$12,730 over a three-year period.

For training using EKS, we assume higher instance utilization and four data scientists can share one instance which is running during business hours. As a result, the number of instances running with EKS are 2 in year 1, 2 in year 2, and 2 in year 3¹³. With the blended cost of instance per hour of \$0.34, the training cost is \$1,414 in year 1, \$1,414 in year 2, and \$1,414 in year 3, totaling \$4,243 over a three-year period.

¹² We have not factored in holidays / vacations when calculating business hours

¹³ Instance count is rounded up to a whole number

For training using Amazon SageMaker, customers only pay for the duration that the training is running aggregated across the year. The total training duration is 260 hours in year 1, 299 hours in year 2, and 344 hours in year 3. With the blended cost of instance per hour of \$0.47, the training cost is \$122 in year 1, \$141 in year 2, and \$162 in year 3, totaling \$424 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the train phase is $\$73 + \$12,730 = \$12,802$. The total infrastructure cost for EKS is $\$73 + \$4,243 = \$4,316$. The total infrastructure cost for Amazon SageMaker is $\$92 + \$424 = \$516$.

For the small scenario, we previously calculated that the total operations costs over a three-year period for EC2 and EKS are \$173,625. For Amazon SageMaker, we previously calculated that total operational cost is \$17,363 over a three-year period. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the train phase for EC2 and EKS over a three-year period is \$57,875. The operational cost for the train phase for Amazon SageMaker is \$5,788.

For the small scenario, we previously calculated that the total security and compliance costs over a three-year period for EC2 and EKS are \$190,000. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the train phase for EC2 and EKS over a three-year period is \$63,333. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the train phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total train cost for EC2 over a three-year period is $\$12,802 + \$57,875 + \$63,333 = \$134,011$. The total train cost for EKS over a three-year period is $\$4,316 + \$57,875 + \$63,333 = \$125,524$. The total train cost for Amazon SageMaker over a three-year period is $\$516 + \$5,788 + \$0 = \$6,304$. The TCO of Amazon SageMaker for a small-sized business with five data scientists for the train phase is 95% lower than EC2 and 95% lower than EKS.

Deploy

The deployment costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$218 and the cost for Amazon SageMaker is \$276. The storage costs are allocated equally among the three phases of ML and the storage cost for the deploy phase over a three-year period for EC2 and EKS is \$73. The storage cost for the deploy phase over a three-year period for Amazon SageMaker is \$92.
- **Network costs:** Over a three-year period, total network cost for EC2 and EKS is \$48 and the cost for Amazon SageMaker is \$67.
- **Compute costs:** The compute cost is calculated as the total duration for which instances are in use for running in a year * cost per instance hour aggregated across a three-year period.

Based on the assumptions for this scenario, 95% of the workload mix is classic ML and the remaining 5% workload mix is deep learning. We are using M5.xlarge for deploying classic ML workloads and P3.2xlarge for deploying deep learning workloads. The cost per instance hour for m5.xlarge and p3.2xlarge in EC2 and EKS is \$0.192 and \$3.06 respectively. Based on the workload mix, the blended instance cost per hour in EC2 and EKS is $(0.95 * \$0.192) + (0.05 * \$3.06) = \$0.34$. The cost per instance hour for ml.m5.xlarge and ml.p3.2xlarge in Amazon SageMaker is \$0.269 and \$4.284 respectively. The blended instance cost per hour in Amazon SageMaker is $(0.95 * \$0.269) + (0.05 * \$4.284) = \$0.47$.

Based on the assumptions for the scenario and a yearly growth rate of 15%, there are 2 models in production in year 1, 3 models in production in year 2, and 4 models in production in year 3¹⁴. We also assume that each model requires one instance for serving and the instances are always available (24 hours in a day, 7 days a week, throughout the year) to serve inferences in real time. The costs are aggregated over a three-year period.

For deploying models using EC2, we assume that each model needs to be deployed on a single instance. As a result, the number of instances running with EC2 are 2 in year 1, 3 in year 2, and 4 in year 3¹⁵. With the blended cost of instance per hour of \$0.34, the deployment cost is \$5,957 in year 1, \$8,935 in year 2, and \$11,914 in year 3, totaling \$26,806 over a three-year period.

For deploying models using EKS, we assume higher instance utilization and that four models can be deployed on a single instance. As a result, the number of instances running with EKS are 1 in year 1, 1 in year 2, and 1 in year 3¹⁶. With the blended cost of instance per hour of \$0.34, the deployment cost is \$2,978 in year 1, \$2,978 in year 2, and \$2,978 in year 3, totaling \$8,935 over a three-year period.

Amazon SageMaker Multi-model endpoints enable you to deploy multiple models on a single endpoint, thereby reducing the number of instances you need available to serve inferences. While the actual number of models you can deploy and run can be in the order of hundreds or thousands, for the TCO analysis, we assume that 4 models can be deployed on a single endpoint. As a result, the number of instances running with Amazon SageMaker are 1 in year 1, 1 in year 2, and 1 in year 3¹⁷. With the blended cost of instance per hour of \$0.47, the deploy cost is \$4,117 in year 1, \$4,117 in year 2, and \$4,117 in year 3, totaling \$12,352 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the deploy phase is $\$73 + \$48 + \$26,806 = \$26,926$. The total infrastructure cost for EKS is $\$73 + \$48 + \$8,935 = \$9,055$. The total infrastructure cost for Amazon SageMaker is $\$92 + \$67 + \$12,352 = \$12,510$.

¹⁴ Number of models are rounded up to a whole number

¹⁵ Number of instances are rounded up to a whole number

¹⁶ Instance count is rounded up to a whole number

¹⁷ Instance count is rounded up to a whole number

For the small scenario, we previously calculated that the total operations costs over a three-year period for EC2 and EKS are \$173,625. For Amazon SageMaker, we previously calculated that total operational costs is \$17,363 over a three-year period. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the deploy phase for EC2 and EKS over a three-year period is \$57,875. The operational cost for the deploy phase for Amazon SageMaker is \$5,788.

For the small scenario, we previously calculated that the total security and compliance costs over a three-year period for EC2 and EKS are \$190,000. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the deploy phase for EC2 and EKS over a three-year period is \$63,333. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the deploy phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total cost of deploy phase for EC2 over a three-year period is $\$26,926 + \$57,875 + \$63,333 = \$148,134$. The total cost of deploy phase for EKS over a three-year period is $\$9,055 + \$57,875 + \$63,333 = \$130,264$. The total cost of deploy phase for Amazon SageMaker over a three-year period is $\$12,510 + \$5,788 + \$0 = \$18,298$. The TCO of Amazon SageMaker for a small-sized business with five data scientists for the deploy phase is 88% lower than EC2 and 86% lower than EKS.

Total costs¹⁸

The total cost for the small scenario are calculated as the sum of the build, train, and deploy costs over a three-year period. The total cost over three-years for EC2 is $\$127,840 + \$134,011 + \$148,134 = \$409,985$. The total cost for EKS over three-years is $\$127,840 + \$125,524 + \$130,264 = \$383,628$. The total cost over for Amazon SageMaker three-years is $\$15,056 + \$6,304 + \$18,298 = \$39,658$. Based on the total costs, Amazon SageMaker has a 90% lower TCO compared to EC2 and a 90% lower TCO compared to EKS.

¹⁸ Some line items may not completely match detailed breakouts due to rounding.

Scenario 2: Medium-sized business

The TCO of Amazon SageMaker for a medium-sized business with 15 data scientists is up to 87% lower than EC2 and up to 85% lower than EKS. The following table presents a summary of the analysis.

3 year TCO Analysis					
Medium Scenario	Amazon SageMaker	Amazon EC2	Amazon EKS	Amazon SageMaker TCO v/s EC2	Amazon SageMaker TCO v/s EKS
TOTAL COSTS (Build + Train + Deploy)	\$103,388	\$815,197	\$700,487	-87%	-85%
Total Build Costs	\$35,011	\$228,657	\$228,657	-85%	-85%
Infrastructure Costs	\$26,736	\$19,241	\$19,241		
Operational Costs	\$8,275	\$82,750	\$82,750		
Security & Compliance Costs	\$ -	\$126,667	\$126,667		
Total Train Costs	\$12,796	\$276,902	\$228,418	-95%	-94%
Infrastructure Costs	\$4,521	\$67,486	\$19,001		
Operational Costs	\$8,275	\$82,750	\$82,750		
Security & Compliance Costs	\$ -	\$126,667	\$126,667		
Total Deploy Costs	\$55,580	\$309,637	\$243,412	-82%	-77%
Infrastructure Costs	\$47,305	\$100,221	\$33,995		
Operational Costs	\$8,275	\$82,750	\$82,750		
Security & Compliance Costs	\$ -	\$126,667	\$126,667		

Build costs

The build costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$1,966 and the cost for Amazon SageMaker is \$2,204. The storage costs are allocated equally among the three phases of ML and the storage cost for the build phase over a three-year period for EC2 and EKS is \$655. The storage cost for the build phase over a three-year period for Amazon SageMaker is \$735.
- **Network costs:** There are no network costs for the build phase.
- **Compute costs:** The compute cost is calculated as the total notebook instances running * hours in a year * cost per hour aggregated across a three-year period. With a 10% yearly

growth rate in the team size, the total notebook instances running for this scenario are 15 in year 1, 17 in year 2, and 19 in year 3¹⁹. For the build phase, each data scientist is running one notebook instance of T3.medium for 24 hours every day of the year. The cost per instance per hour is \$0.0416 in EC2 and \$0.0582 in Amazon SageMaker.

The compute cost for EC2 is \$5,466 in year 1, \$6,195 in year 2, and \$6,924 in year 3, totaling \$18,585 over a three-year period.

The compute cost for EKS is \$5,466 in year 1, \$6,195 in year 2, and \$6,924 in year 3, totaling \$18,585 over a three-year period.

The compute cost for Amazon SageMaker is \$7,647 in year 1, \$8,667 in year 2, and \$9,687 in year 3, totaling \$26,001 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the build phase is $\$655 + \$18,585 = \$19,241$. The infrastructure cost for EKS in the build phase is $\$655 + \$18,585 = \$19,241$. The total infrastructure cost for Amazon SageMaker is $\$735 + \$26,001 = \$26,736$.

For the medium scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$248,250 across all ML phases. For Amazon SageMaker, we previously calculated that total operational cost is \$24,825 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the build phase for EC2 and EKS over a three-year period is \$82,750. The operational cost for the build phase for Amazon SageMaker over a three-year period is \$8,275.

For the medium scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$380,000 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the build phase for EC2 and EKS over a three-year period is \$126,667. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the build phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total build cost for EC2 over a three-year period is $\$19,241 + \$82,750 + \$126,667 = \$228,657$. The total build cost for EKS over a three-year period is $\$19,241 + \$82,750 + \$126,667 = \$228,657$. The total build cost for Amazon SageMaker over a three-year period is $\$26,736 + \$8,275 + \$0 = \$35,011$. The TCO of Amazon SageMaker for a medium-sized business with 15 data scientists for the build phase is 85% lower than EC2 and 85% lower than EKS.

Train

The training costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

¹⁹ Instance count and number of data scientists are rounded up

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$1,966 and the cost for Amazon SageMaker is \$2,204. The storage costs are allocated equally among the three phases of ML and the storage cost for the train phase over a three-year period for EC2 and EKS is \$655. The storage cost for the train phase over a three-year period for Amazon SageMaker is \$735.
- **Network costs:** There are no network costs for the train phase.
- **Compute costs:** The compute cost is calculated as the total duration for which instances are in use for training in a year * cost per instance hour aggregated across a three-year period. Based on the assumptions for this scenario, 85% of the workload mix is classic ML and the remaining 15% workload mix is deep learning. We are using M5.xlarge for training classic ML workloads and P3.2xlarge for training deep learning workloads.

The cost per instance hour for m5.xlarge and p3.2xlarge in EC2 and EKS is \$0.192 and \$3.06 respectively. Based on the workload mix, the blended instance cost per hour in EC2 and EKS is $(0.85 * \$0.192) + (0.15 * \$3.06) = \$0.63$. The cost per instance hour for ml.m5.xlarge and ml.p3.2xlarge in Amazon SageMaker is \$0.269 and \$4.284 respectively. The blended instance cost per hour in Amazon SageMaker is $(0.85 * \$0.269) + (0.15 * \$4.284) = \$0.88$.

For the medium scenario we assume that the data scientists on the team may be training during business hours, i.e. 5 days each week and 8 hours each day for 52 weeks in a year²⁰. Based on the 10% yearly growth rate, the number of data scientists increase from 15 in year 1, to 17 data scientists in year 2 and 19 data scientists in year 3. We assume that the total training duration aggregated across all data scientists is 5 hours per business day in Year 1. With a yearly growth of 10% the total training duration aggregated across all data scientists is 5.5 hours per business day in Year 2, and 6.1 business hours per business day in Year 3²¹.

For training using EC2, we assume that each data scientist needs to provision an instance that is running during business hours. As a result, the number of instances running with EC2 are 15 in year 1, 17 in year 2, and 19 in year 3. With the blended cost of instance per hour of \$0.63, the training cost is \$19,656 in year 1, \$22,277 in year 2, and \$24,898 in year 3, totaling \$66,831 over a three-year period.

For training using EKS, we assume higher instance utilization and four data scientists can share one instance which is running during business hours. As a result, the number of instances running with EKS are 4 in year 1, 5 in year 2, and 5 in year 3²². With the blended cost of instance per hour of \$0.63, the training cost is \$5,242 in year 1, \$6,552 in year 2, and \$6,552 in year 3, totaling \$18,346 over a three-year period.

²⁰ We have not factored in holidays / vacations when calculating business hours

²¹ Some line items may not completely match detailed breakouts due to rounding.

²² Instance count is rounded up to a whole number

For training using Amazon SageMaker, customers only pay for the duration that the training is running aggregated across the year. The total training duration is 1,300 hours in year 1, 1,430 hours in year 2, and 1,573 hours in year 3. With the blended cost of instance per hour of \$0.88, the training cost is \$1,144 in year 1, \$1,258 in year 2, and \$1,384 in year 3, totaling \$3,786 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the train phase is $\$655 + \$66,831 = \$67,486$. The total infrastructure cost for EKS is $\$655 + \$18,346 = \$19,001$. The total infrastructure cost for Amazon SageMaker is $\$735 + \$3,786 = \$4,521$.

For the medium scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$248,250 across all ML phases. For Amazon SageMaker, we previously calculated that total operational cost is \$24,825 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the train phase for EC2 and EKS over a three-year period is \$82,750. The operational cost for the train phase for Amazon SageMaker over a three-year period is \$8,275.

For the medium scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$380,000 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the train phase for EC2 and EKS over a three-year period is \$126,667. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the train phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total train cost for EC2 over a three-year period is $\$67,486 + \$82,750 + \$126,667 = \$276,902$. The total train cost for EKS over a three-year period is $\$19,001 + \$82,750 + \$126,667 = \$228,418$. The total train cost for Amazon SageMaker over a three-year period is $\$4,521 + \$8,275 = \$12,796$. The TCO of Amazon SageMaker for a medium-sized business with 15 data scientists for the train phase is 95% lower than EC2 and 94% lower than EKS.

Deploy

The deployment costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$1,966 and the cost for Amazon SageMaker is \$2,204. The storage costs are allocated equally among the three phases of ML and the storage cost for the train phase over a three-year period for EC2 and EKS is \$655. The storage cost for the train phase over a three-year period for Amazon SageMaker is \$735.
- **Network costs:** Over a three-year period total network cost for EC2 and EKS is \$227 and the cost for Amazon SageMaker is \$318.

- **Compute costs:** The compute cost is calculated as the total duration for which instances are in use for running in a year * cost per instance hour aggregated across a three-year period.

Based on the assumptions for this scenario, 85% of the workload mix is classic ML and the remaining 15% workload mix is deep learning. We are using M5.xlarge for deploying classic ML workloads and P3.2xlarge for deploying deep learning workloads. The cost per instance hour for m5.xlarge and p3.2xlarge in EC2 and EKS is \$0.192 and \$3.06 respectively. Based on the workload mix, the blended instance cost per hour in EC2 and EKS is $(0.85 * \$0.192) + (0.15 * \$3.06) = \$0.63$. The cost per instance hour for ml.m5.xlarge and ml.p3.2xlarge in Amazon SageMaker is \$0.269 and \$4.284 respectively. The blended instance cost per hour in Amazon SageMaker is $(0.85 * \$0.269) + (0.15 * \$4.284) = \$0.88$. Based on the assumptions for the scenario and a yearly growth rate of 10%, there are 5 models in production in year 1, 6 models in production in year 2, and 7 models in production in year 3²³. We also assume that each model requires one instance for serving and the instances are always available (24 hours in a day, 7 days a week, throughout the year) to serve inferences in real time. The costs are aggregated over a three-year period.

For deploying models using EC2, we assume that each model needs to be deployed on a single instance. As a result, the number of instances running with EC2 are 5 in year 1, 6 in year 2, and 7 in year 3²⁴. With the blended cost of instance per hour of \$0.34, the deployment cost is \$27,594 in year 1, \$33,113 in year 2, and \$38,632 in year 3, totaling \$99,338 over a three-year period.

For deploying models using EKS, we assume higher instance utilization and that four models can be deployed on a single instance. As a result, the number of instances running with EKS are 2 in year 1, 2 in year 2, and 2 in year 3²⁵. With the blended cost of instance per hour of \$0.34, the deployment cost is \$11,038 in year 1, \$11,038 in year 2, and \$11,038 in year 3, totaling \$33,113 over a three-year period.

Amazon SageMaker Multi-model endpoints enable you to deploy multiple models on a single endpoint, thereby reducing the number of instances you need available to serve inferences. While the actual number of models you can deploy and run can be in the order of hundreds or thousands, for the TCO analysis, we assume that 4 models can be deployed on a single endpoint. As a result, the number of instances running with Amazon SageMaker are 2 in year 1, 2 in year 2, and 2 in year 3²⁶. With the blended cost of instance per hour of \$0.47, the training cost is \$15,418 in year 1, \$15,418 in year 2, and \$15,418 in year 3, totaling \$46,253 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the deploy phase is $\$655 + \$227 + \$99,338 = \$100,221$.

²³ Number of models are rounded up to a whole number

²⁴ Number of instances are rounded up to a whole number

²⁵ Instance count is rounded up to a whole number

²⁶ Instance count is rounded up to a whole number

The total infrastructure cost for EKS is $\$655 + \$227 + \$33,113 = \$33,995$. The total infrastructure cost for Amazon SageMaker is $\$735 + \$318 + \$46,253 = \$47,305$.

For the medium scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$248,250 across all ML phases. For Amazon SageMaker, we previously calculated that total operational cost is \$24,825 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the deploy phase for EC2 and EKS over a three-year period is \$82,750. The operational cost for the deploy phase for Amazon SageMaker over a three-year period is \$8,275.

For the medium scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$380,000 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the deploy phase for EC2 and EKS over a three-year period is \$126,667. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the deploy phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total cost of deploy phase for EC2 over a three-year period is $\$100,221 + \$82,750 + \$126,667 = \$309,637$. The total cost of deploy phase for EKS over a three-year period is $\$33,995 + \$82,750 + \$126,667 = \$243,412$. The total cost of deploy phase for Amazon SageMaker over a three-year period is $\$47,305 + \$8,275 = \$55,580$. The TCO of Amazon SageMaker for a medium-sized business with 15 data scientists for the deploy phase is 82% lower than EC2 and 77% lower than EKS.

Total costs²⁷

The total cost for the medium scenario are calculated as the sum of the build, train, and deploy costs over a three-year period. The total cost over three-years for EC2 is $\$228,657 + 276,902 + \$309,637 = \$815,197$. The total cost for EKS over three-years is $\$228,657 + \$228,418 + \$243,412 = \$700,487$. The total cost over for Amazon SageMaker three-years is $\$35,011 + \$12,796 + \$55,580 = \$103,388$. Based on the total costs, Amazon SageMaker has an 87% lower TCO compared to EC2 and an 85% lower TCO compared to EKS.

²⁷ Some line items may not completely match detailed breakouts due to rounding.

Scenario 3: Large-sized business

The TCO of Amazon SageMaker for a large-sized business with 50 data scientists is up to 79% lower than EC2 and up to 65% lower than EKS. The following table presents a summary of the analysis.

3 year TCO Analysis					
Large Scenario	Amazon SageMaker	Amazon EC2	Amazon EKS	Amazon SageMaker TCO v/s EC2	Amazon SageMaker TCO v/s EKS
TOTAL COSTS (Build + Train + Deploy)	\$635,918	\$3,060,984	\$1,818,467	-79%	-65%
Total Build Costs	\$105,576	\$492,656	\$492,656	-79%	-79%
Infrastructure Costs	\$88,062	\$64,184	\$64,184		
Operational Costs	\$17,514	\$175,139	\$175,139		
Security & Compliance Costs	\$ -	\$253,333	\$253,333		
Total Train Costs	\$61,888	\$973,788	\$573,721	-94%	-89%
Infrastructure Costs	\$44,375	\$545,316	\$145,248		
Operational Costs	\$17,514	\$175,139	\$175,139		
Security & Compliance Costs	\$ -	\$253,333	\$253,333		
Total Deploy Costs	\$468,454	\$1,594,540	\$752,090	-71%	-38%
Infrastructure Costs	\$450,940	\$1,166,067	\$323,618		
Operational Costs	\$17,514	\$175,139	\$175,139		
Security & Compliance Costs	\$ -	\$253,333	\$253,333		

Build costs

The build costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$18,726 and the cost for Amazon SageMaker is \$20,996. The storage costs are allocated equally among the three phases of ML and the storage cost for the build phase over a three-year period for EC2 and EKS is \$6,242. The storage cost for the build phase over a three-year period for Amazon SageMaker is \$6,999.
- **Network costs:** There are no network costs for the build phase.
- **Compute costs:** The compute cost is calculated as the total notebook instances running * hours in a year * cost per hour aggregated across a three-year period. With a 5% yearly growth rate in the team size, the total notebook instances running for this scenario are 50

in year 1, 53 in year 2, and 55 in year 3²⁸. For the build phase, each data scientist is running one notebook instance of T3.medium for 24 hours every day of the year. The cost per instance per hour is \$0.0416 in EC2 and \$0.0582 in Amazon SageMaker.

The compute cost for EC2 is \$18,221 in year 1, \$19,314 in year 2, and \$20,407 in year 3, totaling \$57,942 over a three-year period.

The compute cost for EKS is \$18,221 in year 1, \$19,314 in year 2, and \$20,407 in year 3, totaling \$57,942 over a three-year period.

The compute cost for Amazon SageMaker is \$25,492 in year 1, \$27,021 in year 2, and \$28,551 in year 3, totaling \$81,063 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the build phase is $\$6,242 + \$57,942 = \$64,184$. The infrastructure cost for EKS in the build phase is $\$6,242 + \$57,942 = \$64,184$. The total infrastructure cost for Amazon SageMaker is $\$6,999 + \$81,063 = \$88,062$

For the large scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$525,417 across all ML phases. For Amazon SageMaker, we previously calculated that total operational cost is \$52,542 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the build phase for EC2 and EKS over a three-year period is \$175,139. The operational cost for the build phase for Amazon SageMaker over a three-year period is \$17,514.

For the large scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$760,000 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the build phase for EC2 and EKS over a three-year period is \$253,333. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the build phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total build cost for EC2 over a three-year period is $\$64,184 + \$175,139 + \$253,333 = \$492,656$. The total build cost for EKS over a three-year period is $\$64,184 + \$175,139 + \$253,333 = \$492,656$. The total build cost for Amazon SageMaker over a three-year period is $\$88,062 + \$17,514 = \$105,576$. The TCO of Amazon SageMaker for a large-sized business with 50 data scientists for the build phase is 79% lower than EC2 and 79% lower than EKS.

Train

The training costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

²⁸ Instance count and number of data scientists are rounded up

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$18,726 and the cost for Amazon SageMaker is \$20,996. The storage costs are allocated equally among the three phases of ML and the storage cost for the train phase over a three-year period for EC2 and EKS is \$6,242. The storage cost for the train phase over a three-year period for Amazon SageMaker is \$6,999.
- **Network costs:** There are no network costs for the train phase.
- **Compute costs:** The compute cost is calculated as the total duration for which instances are in use for training in a year * cost per instance hour aggregated across a three-year period. Based on the assumptions for this scenario, 50% of the workload mix is classic ML and the remaining 50% workload mix is deep learning. We are using M5.xlarge for training classic ML workloads and P3.2xlarge for training deep learning workloads.

The cost per instance hour for m5.xlarge and p3.2xlarge in EC2 and EKS is \$0.192 and \$3.06 respectively. Based on the workload mix, the blended instance cost per hour in EC2 and EKS is $(0.5 * \$0.192) + (0.5 * \$3.06) = \$1.63$. The cost per instance hour for ml.m5.xlarge and ml.p3.2xlarge in Amazon SageMaker is \$0.269 and \$4.284 respectively. The blended instance cost per hour in Amazon SageMaker is $(0.5 * \$0.269) + (0.5 * \$4.284) = \$2.28$.

For the large scenario we assume that the data scientists on the team may be training during business hours, i.e. 5 days each week and 8 hours each day for 52 weeks in a year²⁹. Based on the growth rate, the number of data scientists increase from 50 in year 1, to 53 data scientists in year 2 and 56 data scientists in year 3. We assume that the total training duration aggregated across all data scientists is 20 hours per business day in Year 1. With a yearly growth of 5% the total training duration aggregated across all data scientists is 21 hours per business day in Year 2, and 22.1 business hours per business day in Year 3³⁰.

For training using EC2, we assume that each data scientist needs to provision an instance that is running during business hours. As a result, the number of instances running with EC2 are 50 in year 1, 53 in year 2, and 56 in year 3. With the blended cost of instance per hour of \$1.63, the training cost is \$169,520 in year 1, \$179,691 in year 2, and \$189,862 in year 3, totaling \$539,074 over a three-year period.

For training using EKS, we assume higher instance utilization and four data scientists can share one instance which is running during business hours. As a result, the number of instances running with EKS are 13 in year 1, 14 in year 2, and 14 in year 3³¹. With the blended cost of instance per hour of \$1.63, the training cost is \$44,075 in year 1, \$47,466 in year 2, and \$47,466 in year 3, totaling \$139,006 over a three-year period.

For training using Amazon SageMaker, customers only pay for the duration that the training is running aggregated across the year. The total training duration is 5,200 hours in year 1, 5,460 hours in year 2, and 5,733 hours in year 3. With the blended cost of

²⁹ We have not factored in holidays / vacations when calculating business hours

³⁰ Some line items may not completely match detailed breakouts due to rounding.

³¹ Instance count is rounded up to a whole number

instance per hour of \$2.28, the training cost is \$11,856 in year 1, \$12,449 in year 2, and \$13,071 in year 3, totaling \$37,376 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the train phase is $\$6,242 + \$539,074 = \$545,316$. The total infrastructure cost for EKS is $\$6,242 + \$139,006 = \$145,248$. The total infrastructure cost for Amazon SageMaker is $\$6,999 + \$37,376 = \$44,375$.

For the large scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$525,417 across all ML phases. For Amazon SageMaker, we previously calculated that total operational cost is \$52,542 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the train phase for EC2 and EKS over a three-year period is \$175,139. The operational cost for the train phase for Amazon SageMaker over a three-year period is \$17,514.

For the large scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$760,000 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the train phase for EC2 and EKS over a three-year period is \$253,333. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the train phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total train cost for EC2 over a three-year period is $\$545,316 + \$175,139 + \$253,333 = \$973,788$. The total train cost for EKS over a three-year period is $\$145,248 + \$175,129 + \$253,333 = \$573,721$. The total train cost for Amazon SageMaker over a three-year period is $\$44,375 + \$17,514 = \$61,888$. The TCO of Amazon SageMaker for a large-sized business with 50 data scientists for the train phase is 94% lower than EC2 and 89% lower than EKS.

Deploy

The deployment costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$18,726 and the cost for Amazon SageMaker is \$20,996. The storage costs are allocated equally among the three phases of ML and the storage cost for the train phase over a three-year period for EC2 and EKS is \$6,242. The storage cost for the train phase over a three-year period for Amazon SageMaker is \$6,999.
- **Network costs:** Over a three-year period total network cost for EC2 and EKS is \$3,243 and the cost for Amazon SageMaker is \$4,540.
- **Compute costs:** The compute cost is calculated as the total duration for which instances are in use for running in a year * cost per instance hour aggregated across a three-year period.

Based on the assumptions for this scenario, 50% of the workload mix is classic ML and the remaining 50% workload mix is deep learning. We are using M5.xlarge for deploying classic ML workloads and P3.2xlarge for deploying deep learning workloads. The cost per instance hour for m5.xlarge and p3.2xlarge in EC2 and EKS is \$0.192 and \$3.06 respectively. Based on the workload mix, the blended instance cost per hour in EC2 and EKS is $(0.5 * \$0.192) + (0.5 * \$3.06) = \$1.63$. The cost per instance hour for ml.m5.xlarge and ml.p3.2xlarge in Amazon SageMaker is \$0.269 and \$4.284 respectively. The blended instance cost per hour in Amazon SageMaker is $(0.5 * \$0.269) + (0.5 * \$4.284) = \$2.28$. Based on the assumptions for the scenario and a yearly growth rate of 5%, there are 25 models in production in year 1, 27 models in production in year 2, and 29 models in production in year 3³². We also assume that each model requires one instance for serving and the instances are always available (24 hours in a day, 7 days a week, throughout the year) to serve inferences in real time. The costs are aggregated over a three-year period.

For deploying models using EC2, we assume that each model needs to be deployed on a single instance. As a result, the number of instances running with EC2 are 25 in year 1, 27 in year 2, and 29 in year 3³³. With the blended cost of instance per hour of \$1.63, the deployment cost is \$356,970 in year 1, \$385,528 in year 2, and \$414,085 in year 3, totaling \$1,156,583 over a three-year period.

For deploying models using EKS, we assume higher instance utilization and that four models can be deployed on a single instance. As a result, the number of instances running with EKS are 7 in year 1, 7 in year 2, and 8 in year 3³⁴. With the blended cost of instance per hour of \$1.63, the deployment cost is \$99,952 in year 1, \$99,952 in year 2, and \$114,230 in year 3, totaling \$314,134 over a three-year period.

Amazon SageMaker Multi-model endpoints enable you to deploy multiple models on a single endpoint, thereby reducing the number of instances you need available to serve inferences. While the actual number of models you can deploy and run can be in the order of hundreds or thousands, for the TCO analysis, we assume that 4 models can be deployed on a single endpoint. As a result, the number of instances running with Amazon SageMaker are 7 in year 1, 7 in year 2, and 8 in year 3³⁵. With the blended cost of instance per hour of \$2.28, the training cost is \$139,810 in year 1, \$139,810 in year 2, and \$159,782 in year 3, totaling \$439,402 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the deploy phase is $\$6,242 + \$3,243 + \$1,156,583 = \$1,166,067$. The total infrastructure cost for EKS is $\$6,242 + \$3,243 + \$314,134 = \$323,618$. The total infrastructure cost for Amazon SageMaker is $\$6,999 + \$4,540 + \$439,302 = \$450,940$.

³² Number of models are rounded up to a whole number

³³ Number of instances are rounded up to a whole number

³⁴ Instance count is rounded up to a whole number

³⁵ Instance count is rounded up to a whole number

For the large scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$525,417 across all ML phases. For Amazon SageMaker, we previously calculated that total operational cost is \$52,542 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the deploy phase for EC2 and EKS over a three-year period is \$175,139. The operational cost for the deploy phase for Amazon SageMaker over a three-year period is \$17,514.

For the large scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$760,000 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the deploy phase for EC2 and EKS over a three-year period is \$253,333. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the deploy phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total cost of deploy phase for EC2 over a three-year period is $\$1,166,067 + \$175,139 + \$253,333 = \$1,594,540$. The total cost of deploy phase for EKS over a three-year period is $\$323,618 + \$175,139 + \$253,333 = \$752,090$. The total cost of deploy phase for Amazon SageMaker over a three-year period is $\$450,940 + 17,514 = \$468,454$. The TCO of Amazon SageMaker for a large-sized business with 50 data scientists for the deploy phase is 71% lower than EC2 and 38% lower than EKS.

Total costs³⁶

The total cost for the large scenario are calculated as the sum of the build, train, and deploy costs over a three-year period. The total cost over three-years for EC2 is $\$492,656 + \$973,788 + \$1,594,540 = \$3,060,984$. The total cost for EKS over three-years is $\$492,656 + \$573,721 + \$752,090 = \$1,818,467$. The total cost over for Amazon SageMaker three-years is $\$105,576 + \$61,888 + \$468,454 = \$635,918$. Based on the total costs, Amazon SageMaker has a 79% lower TCO compared to EC2 and a 65% lower TCO compared to EKS.

³⁶ Some line items may not completely match detailed breakouts due to rounding.

Scenario 4: Extra-large-sized business

The TCO of Amazon SageMaker for extra-large-sized business with 250 data scientists is up to 77% lower than EC2 and up to 54% lower than EKS. The following table presents a summary of the analysis.

3 year TCO Analysis					
X-Large Scenario	Amazon SageMaker	Amazon EC2	Amazon EKS	Amazon SageMaker TCO v/s EC2	Amazon SageMaker TCO v/s EKS
TOTAL COSTS (Build + Train + Deploy)	\$2,460,412	\$10,594,181	\$5,342,908	-77%	-54%
Total Build Costs	\$490,786	\$1,374,026	\$1,374,026	-64%	-64%
Infrastructure Costs	\$426,710	\$311,049	\$311,049		
Operational Costs	\$64,076	\$640,755	\$640,755		
Security & Compliance Costs	\$ -	\$422,222	\$422,222		
Total Train Costs	\$298,770	\$3,704,034	\$1,751,164	-92%	-83%
Infrastructure Costs	\$234,695	\$2,641,057	\$688,186		
Operational Costs	\$64,076	\$640,755	\$640,755		
Security & Compliance Costs	\$ -	\$422,222	\$422,222		
Total Deploy Costs	\$1,670,856	\$5,516,120	\$2,217,718	-70%	-25%
Infrastructure Costs	\$1,606,781	\$4,453,143	\$1,154,740		
Operational Costs	\$64,076	\$640,755	\$640,755		
Security & Compliance Costs	\$ -	\$422,222	\$422,222		

Build costs

The build costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$91,346 and the cost for Amazon SageMaker is \$102,418. The storage costs are allocated equally among the three phases of ML and the storage cost for the build phase over a three-year period for EC2 and EKS is \$30,449. The storage cost for the build phase over a three-year period for Amazon SageMaker is \$34,139.
- **Network costs:** There are no network costs for the build phase.
- **Compute costs:** The compute cost is calculated as the total notebook instances running * hours in a year * cost per hour aggregated across a three-year period. With a 2.5% yearly growth rate in the team size, the total notebook instances running for this scenario are

250 in year 1, 256 in year 2, and 263 in year 3³⁷. For the build phase, each data scientist is running one notebook instance of T3.medium for 24 hours every day of the year. The cost per instance per hour is \$0.0416 in EC2 and EKS and \$0.0582 in Amazon SageMaker.

The compute cost for EC2 is \$91,104 in year 1, \$93,655 in year 2, and \$95,841 in year 3, totaling \$280,600 over a three-year period.

The compute cost for EKS is \$91,104 in year 1, \$93,655 in year 2, and \$95,841 in year 3, totaling \$280,600 over a three-year period.

The compute cost for Amazon SageMaker is \$127,458 in year 1, \$131,027 in year 2, and \$134,086 in year 3, totaling \$392,571 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the build phase is \$30,449 + \$280,600 = \$311,049. The infrastructure cost for EKS in the build phase is \$30,449 + \$280,600 = \$311,049. The total infrastructure cost for Amazon SageMaker is \$34,139 + \$392,571 = \$426,710.

For the extra-large scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$1,922,266 across all ML phases. For Amazon SageMaker, we previously calculated that total operational costs is \$192,227 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the build phase for EC2 and EKS over a three-year period is \$640,755. The operational cost for the build phase for Amazon SageMaker over a three-year period is \$64,076.

For the extra-large scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$1,266,667 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the build phase for EC2 and EKS over a three-year period is \$422,222. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the build phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total build cost for EC2 over a three-year period is \$311,049 + \$640,755 + \$422,222 = \$1,374,026. The total build cost for EKS over a three-year period is \$311,049 + \$640,755 + \$422,222 = \$1,374,026. The total build cost for Amazon SageMaker over a three-year period is \$426,710 + \$64,076 = \$490,786. The TCO of Amazon SageMaker for extra-large-sized business with 250 data scientists for the build phase is 64% lower than EC2 and 64% lower than EKS.

Train

The training costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$91,346 and the cost for Amazon SageMaker is \$102,418. The

³⁷ Instance count and number of data scientists are rounded up

storage costs are allocated equally among the three phases of ML and the storage cost for the train phase over a three-year period for EC2 and EKS is \$30,449. The storage cost for the train phase over a three-year period for Amazon SageMaker is \$34,139.

- **Network costs:** There are no network costs for the train phase.
- **Compute costs:** The compute cost is calculated as the total duration for which instances are in use for training in a year * cost per instance hour aggregated across a three-year period. Based on the assumptions for this scenario, 50% of the workload mix is classic ML and the remaining 50% workload mix is deep learning. We are using M5.xlarge for training classic ML workloads and P3.2xlarge for training deep learning workloads.

The cost per instance hour for m5.xlarge and p3.2xlarge in EC2 and EKS is \$0.192 and \$3.06 respectively. Based on the workload mix, the blended instance cost per hour in EC2 and EKS is $(0.5 * \$0.192) + (0.5 * \$3.06) = \$1.63$. The cost per instance hour for ml.m5.xlarge and ml.p3.2xlarge in Amazon SageMaker is \$0.269 and \$4.284 respectively. The blended instance cost per hour in Amazon SageMaker is $(0.5 * \$0.269) + (0.5 * \$4.284) = \$2.28$.

For the large scenario we assume that the data scientists on the team may be training during business hours, i.e. 5 days each week and 8 hours each day for 52 weeks in a year³⁸. Based on the growth rate, the number of data scientists increase from 250 in year 1, to 257 data scientists in year 2 and 263 data scientists in year 3. We assume that the total training duration aggregated across all data scientists is 110 hours per business day in Year 1. With a yearly growth of 2.5% the total training duration aggregated across all data scientists is 112.8 hours per business day in Year 2, and 115.6 business hours per business day in Year 3³⁹.

For training using EC2, we assume that each data scientist needs to provision an instance that is running during business hours. As a result, the number of instances running with EC2 are 250 in year 1, 257 in year 2, and 263 in year 3. With the blended cost of instance per hour of \$1.63, the training cost is \$847,600 in year 1, \$871,333 in year 2, and \$891,675 in year 3, totaling \$2,610,608 over a three-year period.

For training using EKS, we assume higher instance utilization and four data scientists can share one instance which is running during business hours. As a result, the number of instances running with EKS are 63 in year 1, 65 in year 2, and 66 in year 3⁴⁰. With the blended cost of instance per hour of \$1.63, the training cost is \$213,595 in year 1, \$220,376 in year 2, and \$223,766 in year 3, totaling \$657,738 over a three-year period.

For training using Amazon SageMaker, customers only pay for the duration that the training is running aggregated across the year. The total training duration is 28,600 hours in year 1, 29,315 hours in year 2, and 30,048 hours in year 3. With the blended cost of

³⁸ We have not factored in holidays / vacations when calculating business hours

³⁹ Some line items may not completely match detailed breakouts due to rounding.

⁴⁰ Instance count is rounded up to a whole number

instance per hour of \$2.28, the training cost is \$65,208 in year 1, \$66,838 in year 2, and \$68,509 in year 3, totaling \$200,555 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the train phase is $\$30,449 + \$2,610,608 = \$2,641,057$. The total infrastructure cost for EKS is $\$30,449 + \$657,738 = \$688,186$. The total infrastructure cost for Amazon SageMaker is $\$34,139 + \$200,555 = \$234,695$.

For the extra-large scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$1,922,266 across all ML phases. For Amazon SageMaker, we previously calculated that total operational cost is \$192,227 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the train phase for EC2 and EKS over a three-year period is \$640,755. The operational cost for the train phase for Amazon SageMaker over a three-year period is \$64,076.

For the extra-large scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$1,266,667 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the train phase for EC2 and EKS over a three-year period is \$422,222. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the train phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total train cost for EC2 over a three-year period is $\$2,641,057 + \$640,755 + \$422,222 = \$3,704,034$. The total train cost for EKS over a three-year period is $\$688,186 + \$640,755 + \$422,222 = \$1,751,164$. The total train cost for Amazon SageMaker over a three-year period is $\$234,695 + \$64,076 = \$298,770$. The TCO of Amazon SageMaker for extra-large-sized business with 250 data scientists for the train phase is 92% lower than EC2 and 83% lower than EKS.

Deploy

The deployment costs for Amazon SageMaker include the infrastructure, operational, and security and compliance costs over a three-year period.

The infrastructure costs include the storage, network, and compute costs.

- **Storage costs:** We previously calculated that the total storage cost for EC2 and EKS over a three-year period is \$91,346 and the cost for Amazon SageMaker is \$102,418. The storage costs are allocated equally among the three phases of ML and the storage cost for the deploy phase over a three-year period for EC2 and EKS is \$30,449. The storage cost for the deploy phase over a three-year period for Amazon SageMaker is \$34,139.
- **Network costs:** Over a three-year period total network cost for EC2 and EKS is \$10,545 and the cost for Amazon SageMaker is \$14,763.
- **Compute costs:** The compute cost is calculated as the total duration for which instances are in use for running in a year * cost per instance hour aggregated across a three-year period.

Based on the assumptions for this scenario, 50% of the workload mix is classic ML and the remaining 50% workload mix is deep learning. We are using M5.xlarge for deploying classic ML workloads and P3.2xlarge for deploying deep learning workloads. The cost per instance hour for m5.xlarge and p3.2xlarge in EC2 and EKS is \$0.192 and \$3.06 respectively. Based on the workload mix, the blended instance cost per hour in EC2 and EKS is $(0.5 * \$0.192) + (0.5 * \$3.06) = \$1.63$. The cost per instance hour for ml.m5.xlarge and ml.p3.2xlarge in Amazon SageMaker is \$0.269 and \$4.284 respectively. The blended instance cost per hour in Amazon SageMaker is $(0.5 * \$0.269) + (0.5 * \$4.284) = \$2.28$. Based on the assumptions for the scenario and a yearly growth rate of 2.5%, there are 100 models in production in year 1, 103 models in production in year 2, and 106 models in production in year 3⁴¹. We also assume that each model requires one instance for serving and the instances are always available (24 hours in a day, 7 days a week, throughout the year) to serve inferences in real time. The costs are aggregated over a three-year period.

For deploying models using EC2, we assume that each model needs to be deployed on a single instance. As a result, the number of instances running with EC2 are 100 in year 1, 103 in year 2, and 106 in year 3⁴². With the blended cost of instance per hour of \$1.63, the deployment cost is \$1,427,880 year 1, \$1,470,716 in year 2, and \$1,513,553 in year 3, totaling \$4,412,149 over a three-year period.

For deploying models using EKS, we assume higher instance utilization and that four models can be deployed on a single instance. As a result, the number of instances running with EKS are 25 in year 1, 26 in year 2, and 27 in year 3⁴³. With the blended cost of instance per hour of \$1.63, the deployment cost is \$356,970 in year 1, \$371,249 in year 2, and \$385,582 in year 3, totaling \$1,113,746 over a three-year period.

Amazon SageMaker Multi-model endpoints enable you to deploy multiple models on a single endpoint, thereby reducing the number of instances you need available to serve inferences. While the actual number of models you can deploy and run can be in the order of hundreds or thousands, for the TCO analysis, we assume that 4 models can be deployed on a single endpoint. As a result, the number of instances running with Amazon SageMaker are 25 in year 1, 26 in year 2, and 27 in year 3⁴⁴. With the blended cost of instance per hour of \$2.28, the training cost is \$499,320 in year 1, \$519,293 in year 2, and \$539,266 in year 3, totaling \$1,557,878 over a three-year period.

- The total infrastructure cost is calculated as sum of storage, network, and compute costs. The infrastructure cost for EC2 in the deploy phase is $\$30,449 + \$10,545 + \$4,412,149 = \$4,453,143$. The total infrastructure cost for EKS is $\$30,449 + \$10,545 + \$1,113,746 = \$1,154,740$. The total infrastructure cost for Amazon SageMaker is $\$34,139 + \$14,763 + \$1,557,878 = \$1,606,781$.

⁴¹ Number of models are rounded up to a whole number

⁴² Number of instances are rounded up to a whole number

⁴³ Instance count is rounded up to a whole number

⁴⁴ Instance count is rounded up to a whole number

For the extra-large scenario, we previously calculated that the total operational cost for EC2 and EKS over a three-year period is \$1,922,266 across all ML phases. For Amazon SageMaker, we previously calculated that total operational costs is \$192,227 over a three-year period across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The operational cost for the deploy phase for EC2 and EKS over a three-year period is \$640,755. The operational cost for the deploy phase for Amazon SageMaker over a three-year period is \$64,076.

For the extra-large scenario, we previously calculated that the total security and compliance cost over a three-year period for EC2 and EKS is \$1,266,667 across all ML phases. These costs are equally divided and allocated to each of the three ML phases. The security and compliance cost for the deploy phase for EC2 and EKS over a three-year period is \$422,222. Amazon SageMaker provides these security features out of the box so these costs do not apply for Amazon SageMaker.

The total costs for the deploy phase are calculated as the sum of infrastructure, operational, and security and compliance over a three-year period. The total cost of deploy phase for EC2 over a three-year period is $\$4,453,143 + \$640,755 + \$422,222 = \$5,516,120$. The total cost of deploy phase for EKS over a three-year period is $\$1,154,740 + \$640,755 + \$422,222 = \$2,217,718$. The total cost of deploy phase for Amazon SageMaker over a three-year period is $\$1,606,781 + \$64,076 = \$1,670,856$. The TCO of Amazon SageMaker for extra-large-sized business with 250 data scientists for the deploy phase is 70% lower than EC2 and 25% lower than EKS.

Total costs⁴⁵

The total cost for the extra-large scenario are calculated as the sum of the build, train, and deploy costs over a three-year period. The total cost over three-years for EC2 is $\$1,374,026 + \$3,704,034 + \$5,516,120 = \$10,594,181$. The total cost for EKS over three-years is $\$1,374,026 + 1,751,164 + \$2,217,718 = \$5,342,908$. The total cost over for Amazon SageMaker three-years is $\$490,786 + \$298,770 + \$1,670,856 = \$2,460,412$. Based on the total costs, Amazon SageMaker has a 77% lower TCO compared to EC2 and a 54% lower TCO compared to EKS.

Data scientist productivity gains and time savings

In addition to costs savings, customers of all sizes can realize the benefits of Amazon SageMaker, to help significantly improve developer productivity and accelerate ML initiatives.

Data scientists and ML developers are not that easy to find. It's not only expensive to hire the right candidate, but it can also be extremely difficult to retain them due to the high demand for these roles. Then once they are hired, it's not always easy to get them to focus on the most important tasks. That's because the maturity of the ML infrastructure and development processes often limits productivity. Data scientists spend a large portion of their time on manual tasks such as setting up infrastructure and environments, iterating on model versions,

⁴⁵ Some line items may not completely match detailed breakouts due to rounding.

managing model lifecycles, and rewriting code to make it production-ready. Amazon SageMaker makes it easy to build, train, and deploy models at scale. You can write code, launch and track ML experiments, debug models, and monitor model quality—all through the single unified interface, thereby improving productivity. Amazon SageMaker provides pre-built components and fully managed features that boost data scientist productivity by up to 10 times. Below we explore the productivity gains from Amazon SageMaker during each phase of the ML workflow.

Build phase

During the build phase, you can explore data, experiment with various ML frameworks, algorithms, and build models using your data. Amazon SageMaker provides pre-built and fully managed Jupyter Notebooks that need zero setup, are pre-configured with the latest and most popular ML libraries, and are easy to share with colleagues. Data preparation is a key challenge and time-consuming process for data scientists. Amazon SageMaker enables data preparation as well as the ability to run ad hoc analysis, ETL, and Spark jobs on a fully managed and distributed infrastructure. In addition, to reduce the time it takes to build high-quality models, Amazon SageMaker provides AutoML capabilities and pre-built ML algorithms that cover the most common ML use cases in supervised, unsupervised, and reinforcement learning.

Train phase

During the training phase, you can experiment with data processing techniques, various ML algorithms, and model hyperparameters. The number of iterations and training jobs required before successfully deploying a model in production can range from hundreds to thousands. Amazon SageMaker's automatic model tuning capabilities use ML to find the best model based on customer-defined objectives and reduce the time needed to get to high-quality models. In addition, the debugging capabilities in Amazon SageMaker let you detect issues in training jobs early, better analyze deep learning models, and get more insights into the training performance. Finally, as the number of projects, trials, and experiments in the organization increase to thousands, Amazon SageMaker's experiment management capabilities and ML IDE make it easy to organize, track, and manage ML experiments and training jobs across time and organizations.

Deploy phase

During the deployment phase, you can put your ML models in production and make inferences on unseen data. Amazon SageMaker enables one-click model deployment without the need for any code changes. You can deploy ML models that have been trained in Amazon SageMaker or in different environments, and deploy ML models that have been developed using Amazon SageMaker built-in models or custom-built models. You can deploy the same models to make inferences with low latency on real-time data or in batches. In addition, you can deploy new models seamlessly without any impact on availability or loss of performance. Finally, Amazon SageMaker automatically saves inference requests and responses and analyzes the data collected periodically to detect model drift and data quality issues in

production. This continuous monitoring and alerting enables you take preventative actions such as retraining models without building any additional tooling.

Success stories

The following success stories show how Amazon SageMaker customers have reduced costs and improved productivity.



Coinbase uses ML models on Amazon SageMaker to help with fraud prevention, identity verification, and large-scale compliance. **Using Amazon SageMaker, Coinbase reduced model training times from 20 hours to 10 minutes.**



Intuit developed ML models that can pull a year's worth of bank transactions to find deductible business expenses for customers. **Using Amazon SageMaker, Intuit reduced ML deployment time by 90%, from six months to one week.**



Using Amazon SageMaker, NuData Security prevents credit card fraud by analyzing anonymized user data to detect anomalous activity before a fraudulent transaction occurs. **With Amazon SageMaker, NuData reduced ML development time by 60%, simplified their ML architecture by 95%, and worked with a large bank to passively block nearly 100% of fraudulent attempt traffic within the bank's consumer friction tolerance.**



Using Amazon SageMaker, Voodoo can decide in real time which ad to show to their players and invoke their endpoint over 100 million times by over 30 million users daily, representing close to a billion predictions per day. **With AWS machine learning, Voodoo put an accurate model into production in less than a week, supported by a small team, and has built on top of it continuously as their team and business grow.**



Using TensorFlow on Amazon SageMaker, Siemens Financial Services developed an NLP model to extract critical information to accelerate investment due diligence, **reducing time to summarize diligence documents from 12 hours to 30 seconds.**



Celgene uses Apache MXNet on Amazon SageMaker for toxicology prediction to analyze biological impacts of potential drugs virtually, without putting patients at risk. **A model that previously took two months to train can now be trained in four hours.**



ADP uses AWS machine learning, including Amazon SageMaker, to quickly identify workforce patterns and predict outcomes before they happen, for example employee turnover or the impact of an increase in compensation. **ADP reduced the time to deploy machine learning models from 2 weeks to just 1 day.**

Conclusion

Amazon SageMaker is a fully managed ML service that lets you build, train, tune, and deploy models at scale. It eliminates the overhead costs of provisioning hardware and systems software, managing infrastructure and operations, and building security and compliance for ML workloads. As a result, you can focus on your business problems, develop with new modeling techniques, and take ML projects from idea to production faster. The total cost of ownership of Amazon SageMaker over a three-year horizon can be 54% lower compared to other cloud-based ML options such as self-managed EC2 and managed Kubernetes (K8s) on AWS like EKS. In addition to the lower TCO, Amazon SageMaker's fully managed and integrated features lets you put ML ideas into production faster and improve data scientist productivity by up to 10 times.