# Cost Optimization: EC2 Right Sizing

## AWS Implementation Guide

*Chris Han Sijie*

*Fernando Teodoro de Lima*

*Bryan Miller*

November 2016

*Last updated: December 2016 (see revisions)*

## Contents

## About This Guide

This implementation guide discusses architectural considerations and configuration steps for deploying the Cost Optimization: EC2 Right Sizing solution on the Amazon Web Services (AWS) Cloud. It includes links to AWS CloudFormation templates that launch, configure, and run the AWS compute, network, storage, and other services required to deploy this solution on AWS, using AWS best practices for security and availability.

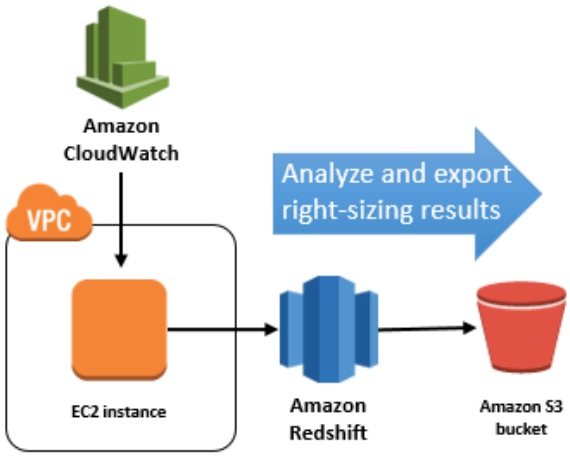
The guide is intended for IT infrastructure architects, administrators, and DevOps professionals who have practical experience architecting on the AWS Cloud.

# Overview

Amazon Web Services (AWS) enables customers to generate reports to gain insight into service usage and costs as they deploy and operate cloud architectures. This includes detailed reports and metrics, which customers can access in the AWS Management Console, Amazon CloudWatch, and AWS Trusted Advisor. Organizations can analyze this information to better understand how to leverage the elasticity of the AWS Cloud to optimize their costs yet still meet their performance and capacity requirements.[1]

Amazon Elastic Compute Cloud (Amazon EC2) provides a wide selection of instance types and sizes optimized to fit different use cases, giving customers the flexibility to choose the appropriate mix of resources for their applications. Amazon EC2 generates detailed usage data to help determine the *right sizing* of their Amazon EC2 resources, which indicates whether they are using the most cost effective instance to meet the technical requirements of the given workload.

AWS offers the Cost Optimization: EC2 Right Sizing (EC2 Right Sizing) solution, an automated reference deployment that analyzes of two weeks of Amazon EC2 utilization data and provides detailed right-sizing recommendations to meet the current demand while reducing the overall cost to run the workload. The solution leverages AWS managed services, enabling deployment in a highly available and affordable way.




This guide provides infrastructure and configuration information for planning and deploying the EC2 Right Sizing solution. The following sections assume basic knowledge of Amazon EC2, Amazon Redshift, and Amazon CloudWatch. It is also helpful to have working knowledge of comma-separated values (.csv) files and Microsoft Excel.

[1] For more information on cost optimization, see the AWS website.

## Cost

You are responsible for the cost of the AWS services used while running this reference deployment. As of the date of publication, the cost for running EC2 Right Sizing with this solution's default settings in the US East (N. Virginia) Region is approximately **$0.65 per hour**. This reflects Amazon Redshift and Amazon EC2 charges.

You will also incur variable charges from Amazon Simple Storage Service (Amazon S3) and Amazon CloudWatch. For full details, see the pricing webpage for each AWS service you will be using in this solution.

Note that this solution is intended for temporary use; after you deploy the AWS CloudFormation stack and save the results, you have the option to automatically delete compute resources to stop incurring unnecessary costs. You can deploy this solution as often as necessary to reanalyze your Amazon EC2 sizing.

## Architecture Overview

Deploying this solution with the **default parameters** builds the following environment in the AWS Cloud.

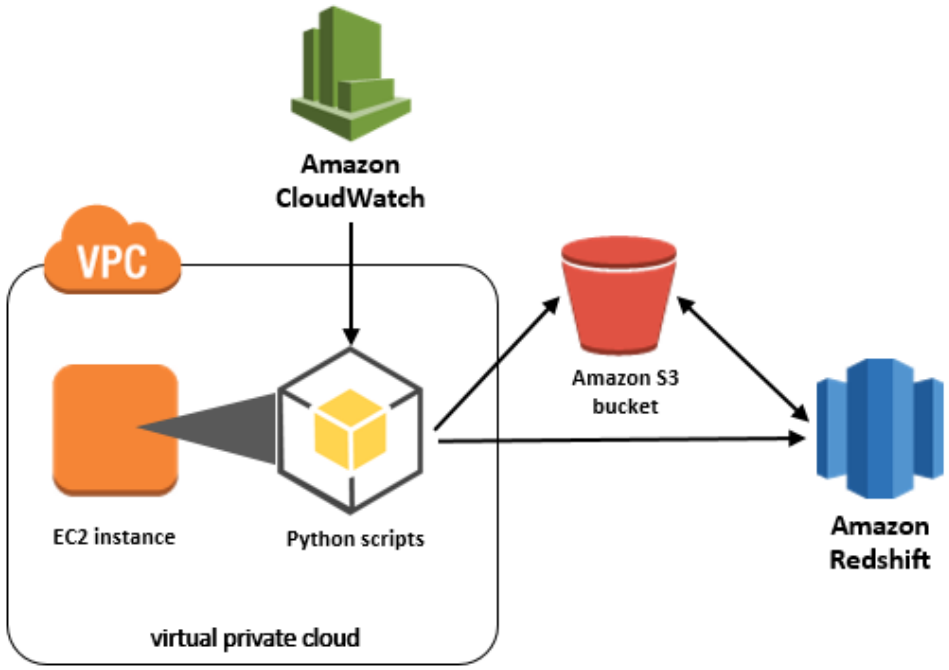



**Figure 1: EC2 Right Sizing solution architecture on AWS**

This solution uses AWS CloudFormation to deploy AWS resources and Python code to provide a right-sizing analysis for all Amazon EC2 instances in a customer account. The AWS CloudFormation template launches a two-node Amazon Redshift cluster, using dc1.large node types. The solution also deploys an Amazon EC2 instance in an Amazon Virtual Private Cloud (Amazon VPC) network. The instance hosts a sequence of Python scripts that collect utilization data from Amazon CloudWatch and then run a custom

query in a temporary Amazon Redshift cluster to produce the right-sizing analysis. Both the raw CloudWatch data and the analysis (CSV format) are stored in an Amazon S3 bucket. Users have the option to automatically terminate the Amazon EC2 instance and Amazon Redshift cluster after the analysis is delivered to reduce ongoing cost. After downloading the analysis from Amazon S3, users can then manually delete the AWS CloudFormation stack.

As mentioned in the Cost section, this solution is designed for temporary deployment in a customer's account. Customers can deploy this solution every two weeks for continuous monitoring, or whenever they want to analyze their Amazon EC2 sizing.

# Implementation Considerations

## Right-Sizing Recommendations

This solution offers recommendations as a starting point to help identify incorrectly sized Amazon EC2 instances. The resulting .csv file provides an analysis of each instance's provisioned size and utilization to help customers determine the most appropriate instance type for their workloads. For detailed information on how the results are calculated, see Appendix A: Right Sizing Logic.

This solution provides prescriptive instance type recommendations. Before incorporating these changes into your overall capacity management processes, we recommend that you test the proposed Amazon EC2 instances to ensure they are properly sized to fulfill their expected role.

## Data Limits

Note that the CloudWatch metrics used in the analysis reflect the last two weeks of Amazon EC2 usage data. Therefore, be aware of the impact of seasonal or business cycles on the metrics and results.

The EC2 Right Sizing solution pulls utilization metrics for instances that are running or stopped (at the time of solution deployment) in order to provide recommendations for existing resources. This solution does not analyze usage or provide recommendations for instances that were terminated manually, as part of an Auto Scaling group, or for any other reason.

For customers who use Consolidated Billing to manage payment for multiple accounts, this solution will analyze Amazon EC2 instances only for the account in which it is deployed (whether that is the payer account or a different linked account).

## Regional Deployments

This EC2 Right Sizing solution uses an AWS Lambda function to configure the stack resources, therefore you must deploy the solution in an AWS Region that supports AWS

Lambda.[2] Once deployed, the solution will automatically analyze usage data for all Amazon EC2 instances in all AWS Regions of a customer's account.[3]

# AWS CloudFormation Template

This solution uses AWS CloudFormation to automate the deployment of the EC2 Right Sizing solution on the AWS Cloud. It includes the following AWS CloudFormation template, which you can download before deployment:

View template

**cost-optimization-ec2-right-sizing.template:** Use this template to launch the Cost Optimization: EC2 Right Sizing solution and all associated components, as described in the previous section. You can also customize the template based on your specific needs.

# Automated Deployment

Before you launch the automated deployment, please review the architecture, configuration and other considerations discussed in this guide. Follow the step-by-step instructions in this section to configure and deploy the EC2 Right Sizing solution into your account.

**Time to deploy:** Approximately 25 minutes.

## What We'll Cover

The procedure for deploying this architecture on AWS consists of the following steps. For detailed instructions, follow the links for each step.

Step 1. Launch the Stack

- Launch the AWS CloudFormation template into your AWS account.
- Enter values for required parameters: **Stack name**, **SSH Key**, **Access CIDR Block**
- Review the other template parameters, and adjust if necessary.

Step 2. Download the Right-Sizing Results

- Download the .csv file from Amazon S3.
- Delete the solution's AWS CloudFormation stack.

[2] For the most current AWS Lambda availability by region, see https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/

[3] At the time of publication, this does not include AWS GovCloud (US), Asia Pacific (Singapore) Region, or China (Beijing) Region.

# Step 1. Launch the Stack

This automated AWS CloudFormation template deploys the EC2 Right Sizing solution on the AWS Cloud.

> **Note**: You are responsible for the cost of the AWS services used while running this solution. See the Cost section for more details. For full details, see the pricing webpage for each AWS service you will be using in this solution.

1. Log in to the AWS Management Console and click the button to the right to launch the *cost-optimization-ec2-right-sizing* AWS CloudFormation template.
   You can also download the template as a starting point for your own implementation.


2. The template is launched in the US East (N. Virginia) Region by default. To launch the EC2 Right Sizing solution in a different AWS Region, use the region selector in the console navigation bar.

> **Note**: This solution uses an AWS Lambda function to configure the stack resources. AWS Lambda is currently available in specific AWS Regions only, therefore you must launch this solution an AWS Region where the service is available. [4]

3. On the **Select Template** page, verify that you selected the correct template and choose **Next**.
4. On the **Specify Details** page, assign a name to your EC2 Right Sizing stack.
5. Under **Parameters**, review the parameters for the template and modify them as necessary. This solution uses the following default values.

| Parameter | Default | Description |
|---|---|---|
| **SSH Key** | <Requires input> | Public and private key pair, which allows you to connect securely to the Amazon EC2 instance. When you created an AWS account, this is the key pair you created in your preferred AWS Region. |
| **Access CIDR Block** | <Requires input> | This IP address range will have access to the EC2 instance. |
| **Terminate Resources** | `Yes` | Choose `Yes` to automatically terminate the Redshift cluster and EC2 instance once the results have been stored in the S3 bucket. Choose `No` to keep these |

[4] For the most current service availability by AWS Region, see https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/

| Parameter | Default | Description |
|---|---|---|
| | | resources running (you can manually delete them with the AWS CloudFormation stack at a later time).<br><br>**Note:** This option enables you to terminate resources as soon as possible to avoid incurring unnecessary costs. To delete all solution resources (VPC, S3 bucket) you must delete the AWS CloudFormation stack, which is discussed in Step 2. Download the Right-Sizing Results. |
| **Send Anonymous Usage Data** | `Yes` | Send anonymous data to AWS to help us understand solution usage and related cost savings across our customer base as a whole. To opt out of this feature, choose `No`.<br><br>For more information, see Appendix B. |

6. Choose **Next**.
7. On the **Options** page, you can specify tags (key-value pairs) for resources in your stack and set additional options, and then choose **Next**.
8. On the **Review** page, review and confirm the settings. Be sure to check the box acknowledging that the template will create AWS Identity and Access Management (IAM) resources.
9. Choose **Create** to deploy the stack.

   You can view the status of the stack in the AWS CloudFormation console in the **Status** column. After all stack resources have successfully launched, you will see the message CREATE_COMPLETE. This can take 20 or more minutes depending on the number of resources in your account.
10. To see details for the stack resources, choose the **Outputs** tab. The following table describes each of these outputs in more detail.

| Key | Description |
|---|---|
| **BucketName** | Amazon S3 bucket created to hold CloudWatch metrics and the right-sizing results |
| **ClusterName** | Amazon Redshift cluster created by the solution |
| **ClusterEndpoint** | Endpoint of the Amazon Redshift cluster created by the solution |

**Note:** This solution deploys an AWS Lambda function, `solution-helper`, which runs only during initial configuration or when resources are updated or deleted. You will see the `solution-helper` function in the AWS Lambda console, which is necessary to manage associated resources for as long as the solution is running.

## Step 2. Download the Right-Sizing Results

After the EC2 Right Sizing stack launch completes, download the .csv file that contains the right-sizing analysis for your Amazon EC2 resources. See Appendix A for information on how recommendations are calculated.

1. In the AWS CloudFormation console, in the **Outputs** tab, note the name of the Amazon S3 bucket (**BucketName** output) that the solution created.
2. Open the Amazon S3 console and navigate to the applicable bucket.
3. The bucket should contain the right-sizing results file `result_rightsizingXXXXXXXXX.csv`. Download this file to a different location.

   Note that the bucket contains other files that you can download: `YYYY-MM-DD-before336hour-with60min.csv.gz`, which is the raw CloudWatch metrics data, and `ec2pricelist.csv`, which is the latest Amazon EC2 pricing information.
4. Review the right-sizing results file. It includes Amazon EC2 instance data such as AWS Region, current instance type, instance ID, resource tags, maximum CPU usage and IOPS, suggested instance type, and estimated monthly savings.

   **Important:** This solution provides prescriptive instance type recommendations. Before incorporating these changes into your overall capacity management processes, we recommend that you test the proposed Amazon EC2 instances to ensure they are properly sized to fulfill their expected role.
5. After you verify that your downloaded files are valid, delete all files in the Amazon S3 bucket.

   **Note:** You must delete all objects in the Amazon S3 bucket to successfully delete the AWS CloudFormation stack.
6. In the AWS CloudFormation console, delete the stack to delete all solution-related resources.

   **Note:** If you chose not to terminate Amazon Redshift and Amazon EC2 resources automatically in the previous procedure, they will be terminated when you delete the stack.

| region | InstanceType | vCPU | Memory | Storage | NetworkPerformance | OldRate | InstanceId | MaxCPU | MaxIOPS | MaxNetwork | InstanceTag | NewInstanceType | NewRate | CostSavedPerMonth |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| USE1 | m4.xlarge | 4 | 16 GiB | EBS only | High | 0.239 | i-1d1d1d1d | 10.21 | 0.00E+00 | 0.096083323 | name:example1 \| awss | m3.medium | 0.067 | 123.84 |
| USE1 | m4.xlarge | 4 | 16 GiB | EBS only | High | 0.239 | i-7i7i7i7i | 16.96 | 0.00E+00 | 0.539735158 | tagname2:example2 \| ' | m3.medium | 0.067 | 123.84 |
| USW2 | m4.xlarge | 4 | 16 GiB | EBS only | High | 0.239 | i-4a4a4a4a | 7.08 | 0.00E+00 | 3.292332967 | anotherTagKey:exampl | m3.medium | 0.067 | 123.84 |
| USE1 | m3.xlarge | 4 | 15 GiB | 2 x 40 SSD | High | 0.266 | i-1p1p1p1p | 7.67 | 0.00E+00 | 11.90900714 | tagname2:example2 \| ! | m3.medium | 0.067 | 143.28 |
| USE1 | c4.4xlarge | 16 | 30 GiB | EBS only | High | 0.838 | i-5a5a5a5a | 39.46 | 0.00E+00 | 0.051218414 | anotherTagKey:exampl | c4.2xlarge | 0.419 | 301.68 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | 1-4e4e4e4e | 9.35 | 9392.35 | 6.340506236 | awscloudformationstac | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1g1g1g1g | 9.28 | 9385.416667 | 4.512665558 | anotherTagKey:exampl | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1j1j1j1j | 9.72 | 9395.183333 | 4.731430054 | tagname4:example4 \| ( | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1m1m1m1 | 9.49 | 9383.3 | 4.484388224 | tagname2:example2 \| ( | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-2b2b2b2b | 9.82 | 9412.266667 | 4.495421855 | tagname2:example2 \| ( | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1o1o1o1o | 9.55 | 9399.533333 | 4.596212896 | awscloudformationlogi | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-2g2g2g2g | 9.35 | 9404.933333 | 4.733863831 | anotherTagKey:exampl | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | 1-7e7e7e7e | 9.33 | 9386.433333 | 6.104599508 | tagname3:example3 \| ( | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-2e2e2e2e | 9.66 | 9393.416667 | 9.122511546 | tagname2:example2 \| ' | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-2i2i2i2i | 9.29 | 9397.05 | 10.9415048 | awscloudformationlogi | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1n1n1n1n | 9.41 | 9384.75 | 4.591510646 | awscloudformationlogi | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-5i5i5i5i | 9.51 | 9392.233333 | 5.242523193 | tagname3:example3 \| ( | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1e1e1e1e | 9.53 | 9407.683333 | 13.70971756 | awscloudformationlogi | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | 1-2o2o2o2o | 9.46 | 9395.083333 | 4.645130666 | TerminateScript:Age 27 | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1c1c1c1c | 9.44 | 9407.416667 | 6.662218857 | anotherTagKey:exampl | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-2a2a2a2a | 9.86 | 9394.916667 | 4.971524556 | awscloudformationlogi | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-5i5i5i5i | 9.53 | 9388.5 | 4.536758169 | awscloudformationstac | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-2d2d2d2d | 9.8 | 9386.933333 | 23.51325124 | awscloudformationstac | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-4i4i4i4i | 9.44 | 9396.25 | 5.566611481 | awscloudformationlogi | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1a1a1a1a | 9.28 | 9417.15 | 36.13259354 | anotherTagKey:exampl | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-6i6i6i6i | 9.6 | 9402.6 | 4.838091914 | tagname2:example2 \| ( | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-1b1b1b1b | 9.4 | 9394.633333 | 4.782803345 | anotherTagKey:exampl | c4.8xlarge | 1.675 | 3704.4 |
| USE1 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | i-5i5i5i5i | 9.58 | 9397.816667 | 4.679567464 | TerminateScript:Age 27 | c4.8xlarge | 1.675 | 3704.4 |
| USW2 | i2.8xlarge | 32 | 244 GiB | 8 x 800 SSD | 10 Gigabit | 6.82 | | 2.76 | 15.46666667 | 4.839845149 | TerminateScript:Age 27 | c4.8xlarge | 1.675 | 3704.4 |
| Total | | | | | | | | | | | | | | 89722.08 |

**Figure 2: Example right-sizing analysis**

# Security

The AWS Cloud provides a scalable, highly reliable platform that helps customers deploy applications and data quickly and securely.

When you build systems on AWS infrastructure, security responsibilities are shared between you and AWS. This shared model can reduce your operational burden as AWS operates, manages, and controls the components from the host operating system and virtualization layer down to the physical security of the facilities in which the services operate. In turn, you assume responsibility and management of the guest operating system (including updates and security patches), other associated applications, as well as the configuration of the AWS-provided security group firewall. For more information about security on AWS, visit the AWS Security Center.

## Security Groups

The security groups created in this solution are designed to control and isolate network traffic between the EC2 instance and Amazon Redshift. We recommend that you review the security groups and further restrict access as needed once the deployment is up and running.

## Additional Security Settings

The Redshift cluster requires that the user provide a username and password pair to provide control over access to the Amazon Redshift cluster.

# Additional Resources

## AWS webpages and documentation

- AWS Cost Optimization
- Cost Optimization Monitor
- AWS CloudFormation
- Amazon Redshift
- Amazon EC2 user guide for Linux instances
- AWS SDK for Python (Boto)
- Amazon VPC
- Amazon S3
- AWS IAM

# Appendix A: Right Sizing Logic

After the Amazon CloudWatch metrics are uploaded to Amazon Redshift, the solution submits a query using the following logic.

1. Search for all Amazon EC2 instances with a max CPU utilization lower than 50%.
2. For each instance returned, use the following conditions to determine a more cost-effective replacement:
   - If the instance type has 10 Gigabit network connectivity, the new instance type must also include 10 Gigabit network connectivity.
   - If the instance store (local storage volumes, not Amazon Elastic Block Store volumes) supports more than 3,000 IOPS, the new instance type must meet the following requirements:
     - ✓ Equal or increased I/O performance
     - ✓ CPU performance meets or exceeds the original instance's max CPU utilization
     - ✓ Equal or increased memory
     - ✓ Equal or increased network throughput
     - ✓ Lowest hourly cost among all instance types that meet the above four requirements
   - If the instance store supports less than 3,000 IOPS, the new instance type must meet the following requirements:
     - ✓ CPU performance meets or exceeds the original instance's max CPU utilization
     - ✓ Equal or increased memory
     - ✓ Equal or increased network throughput
     - ✓ Lowest hourly cost among all instance types that meet the above three requirements
3. After identifying an appropriate replacement for each instance, calculate the estimated monthly cost savings and export all relevant data as a .csv file.

# Appendix B: Collection of Anonymous Data

This solution includes an option to send anonymous usage data to AWS. We use this data to better understand how customers use this solution to improve the services and products that we offer. When enabled, the following information is collected and sent to AWS:

- **Solution ID:** The AWS solution identifier
- **Unique ID (UUID):** Randomly generated, unique identifier for each EC2 Right Sizing solution deployment
- **Timestamp:** Data-collection timestamp
- **Cluster Size:** Size of the Amazon Redshift cluster the solution will deploy

Note that AWS will own the data gathered via this survey. Data collection will be subject to the AWS Privacy Policy. To opt out of this feature, set the **Send Anonymous Usage Data** parameter to `No`.

# Send Us Feedback

We welcome your questions and comments. Please post your feedback on the AWS Solutions Discussion Forum.

You can visit our GitHub repository to download the templates and scripts for this solution, and to share your customizations with others.

# Document Revisions

| Date | Change | In sections |
|---|---|---|
| **November 2016** | Initial publication | - |
| **December 2016** | Correction to template download links | AWS CloudFormation Template, Step 1. Launch the Stack |



**Notices**

This document is provided for informational purposes only. It represents AWS's current product offerings and practices as of the date of issue of this document, which are subject to change without notice. Customers are responsible for making their own independent assessment of the information in this document and any use of AWS's products or services, each of which is provided "as is" without warranty of any kind, whether express or implied. This document does not create any warranties, representations, contractual commitments, conditions or assurances from AWS, its affiliates, suppliers or licensors. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

The Cost Optimization: EC2 Right Sizing solution is licensed under the terms of the Amazon Software License available at https://aws.amazon.com/asl/.