Auto Scaling

Quick Start

Quick Start

Activate and authorize the Auto Scaling service

Auto Scaling is a type of regional resources. For more information, see regions and zones. Before using Auto Scaling, you must activate the service. To create clusters or on-demand execution plans, you must grant the AliyunESSDefaultRole role to Auto Scaling, so that it can call ECS, VPC, SLB, and other relevant services. For more information about roles and permissions, see what is RAM.

Authorization steps

Log on to the **Auto Scaling console**.

Click **Authorize** to go to the RAM console and grant authorization.

Select **AliyunESSDefaultRole** and click **Agree to Authorize**.

Return to the Auto Scaling console and refresh the page.

Next step

After completing authorization, you can use the Auto Scaling service. Next, you can create your first scaling group in the desired region (such as **China (Shanghai)**). For more information, see **create a scaling group**.

Permission list

By default, the role AliyunESSDefaultRole allows Auto Scaling to call the following Alibaba Cloud resources for you:

ECS permissions

Permission name	Permission description
ecs:RunInstances	Create one or more ECS instances as needed.
ecs:CreateInstance	Create ECS instances.
ecs:StartInstance	Start ECS instances.
ecs:AllocatePublicIpAddress	Allocate public IP addresses to ECS instances.
ecs:StopInstance	Stop ECS instances.
ecs:DeleteInstance	Delete ECS instances.
ecs:DescribeInstances	Query ECS instance lists.
ecs:DescribeInstanceAttribute	Query ECS instance attributes.
ecs:ModifyInstanceAttribute	Modify ECS instance attributes.
ecs:DescribeSecurityGroupAttribute	Query security group attributes.
ecs:DescribeSnapshots	Query snapshot lists.
ecs:DescribeKeyPairs	Query key pair lists.

SLB permissions

Permission name	Permission description
slb:DescribeLoadBalancerAttribute	Query SLB instance information.
slb:RemoveBackendServers	Delete backend servers from a SLB instance.
slb:DescribeHealthStatus	Check the health of the backend servers of a SLB instance.
slb:AddBackendServers	Add backend servers to a SLB instance.
slb:SetBackendServers	Configure backend server weights.

RDS permissions

Permission name	Permission description
rds:ModifySecurityIps	Modify the IP address whitelist of an RDS

	instance.
rds:DescribeDBInstanceAttribute	View RDS instance details.
rds:DescribeTaskInfo	Query RDS task information.
rds:DescribeDBInstanceIPArrayList	View the IP address whitelist of an RDS instance.

VPC permissions

Permission name	Permission description
vpc:DescribeVpcs	Query VPC lists.
vpc:DescribeVSwitches	Query VSwitch lists.

MNS permissions

Permission name	Permission description
mns:ListTopic	List topic lists under an account.
mns:ListQueue	List queue lists under an account.
mns:SendMessage	Send messages.
mns:PublishMessage	Publish messages.

The complete policy list for the role AliyunESSDefaultRole is as follows:

```
{
"Version": "1",
"Statement": [
{
"Action": [
"ecs:CreateInstance",
"ecs:RunInstances",
"ecs:StartInstance",
"ecs:AllocatePublicIpAddress",
"ecs:StopInstance",
"ecs:DeleteInstance",
"ecs:DescribeInstances",
"ecs:DescribeInstanceAttribute",
"ecs:ModifyInstanceAttribute",
"ecs:DescribeSecurityGroupAttribute",
"ecs:DescribeImages",
"ecs:DescribeSnapshots",
"ecs:DescribeKeyPairs",
"slb:DescribeLoadBalancerAttribute",
"slb:RemoveBackendServers",
"slb:DescribeHealthStatus",
```

```
"slb:AddBackendServers",
"slb:SetBackendServers",
"rds:ModifySecurityIps",
"rds:DescribeDBInstanceAttribute",
"rds:DescribeTaskInfo",
"rds:DescribeDBInstanceIPArrayList"
],
"Resource": "*",
"Effect": "Allow"
},
{
"Action": [
"vpc:DescribeVpcs",
"vpc:DescribeVSwitches"
],
"Resource": "*",
"Effect": "Allow"
},
{
"Action": [
"mns:ListTopic",
"mns:ListQueue",
"mns:SendMessage",
"mns:PublishMessage"
],
"Resource": "*",
"Effect": "Allow"
},
{
"Action": "ram:PassRole",
"Resource": "*",
"Effect": "Allow",
"Condition": {
"StringEquals": {
"acs:Service": "ecs.aliyuncs.com"
}
}
}
]
}
```

Create a simple scaling solution

It takes 2 main steps as follows to create a simple scaling solution.
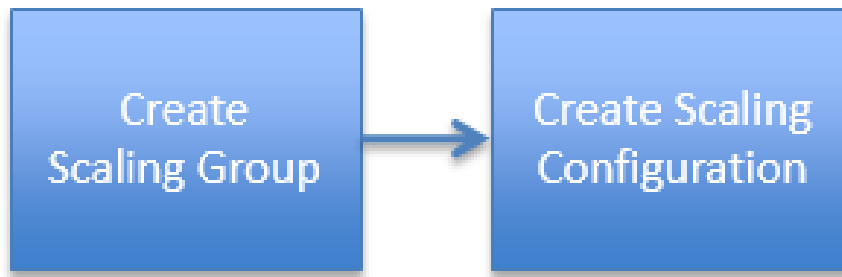
Log on to the **Auto Scaling console**, click **Overall solution creation and management** on the navigation pane.

Create a scaling group

A scaling group includes ECS instances that can be applied to the same scenario. You can define the minimum and maximum number of ECS instances in the group, and the associated Server Load Balancer instances and RDS instances.

Enter a value or make a selection for each option listed.

For the **Minimum Number of Instances Allowed for Scaling**, Auto Scaling starts to create this number of ECS instances in the scaling group.

For **Server Load Balancer**, all the ports listened on by the specified SLB must be enabled health check.

Click **Submit**.

Click **Create scaling configuration**.

Create scaling configuration

You can specify the specifications of the ECS instances used for scaling.

Enter a value or make a selection for each option listed.

For **Image Type**, select **Custom Image** if you need features such as the automated startup of Web servers, the automated downloading of code and scripts.

Click **Preview**.

Preview

The **Preview** page shows the configuration of the previous steps, and the cost of this solution. You are only charged for using the ECS instances, Auto Scaling service is free.

Click **Complete**, you can see whether the scaling group you specified is created.

> **Note**: The scaling group starts by creating the minimum number of instances you specify, and automated adds these instances into the SLB service, and the IPs of these instances into the whitelist of the RDS.