WEBVTT

00:00:00.000 --> 00:00:06.000

Добрый день, уважаемые студенты! Мы с вами начинаем разбор вопроса Practice Test-а.

00:00:06.000 --> 00:00:17.000

Согласно вопросу, компания хостит свое приложение в нескольких availability зонах в регионе us-west-1.

00:00:17.000 --> 00:00:25.000

Используются все три availability зоны. Также используется Auto Scaling group в связке с Application Load Balancer.

00:00:25.000 --> 00:00:35.000

На текущий момент в availability зонах A, B и C соответственно 10, 8 и 7 запущенных инстансов.

00:00:35.000 --> 00:00:46.000

Так как трафик на приложение потихоньку уменьшается, то планируется запустить scale-in операцию.

00:00:46.000 --> 00:01:00.000

Вопрос, каким образом Auto Scaling group будет определять, какой из инстансов нужно терминейтнуть в первую очередь.

00:01:00.000 --> 00:01:04.000

Нам необходимо из предложенных вариантов выбрать 3 правильных.

00:01:04.000 --> 00:01:12.000

На этом слайде вы видите варианты ответов. Также видите, что второй, четвертый и шестой являются верными.

00:01:12.000 --> 00:01:23.000

Здесь я бы хотел обратить ваше внимание, что все варианты ответов содержат противоположный вариант.

00:01:23.000 --> 00:01:28.000

То есть они парные. Это первый и четвертый.

00:01:28.000 --> 00:01:37.000

Далее, это второй и пятый, и третий с шестым.

00:01:37.000 --> 00:01:46.000

Таким образом мы можем понять, что как минимум один из каждой из пар является верным ответом.

00:01:46.000 --> 00:01:53.000

Для того, чтобы ответить на вопрос, давайте посмотрим на объяснение.

00:01:53.000 --> 00:02:00.000

В вопросе говорится, что стоит default termination policy.

00:02:00.000 --> 00:02:12.000

Идея ее в том, что Auto Scaling group стремится распределить количество инстансов равномерно по всем availability зонам.

00:02:12.000 --> 00:02:23.000

Исходя из этого, она старается ответить на 4 вопроса и соответственно принимает решение по отключению того или иного инстанса.

00:02:23.000 --> 00:02:26.000

Давайте посмотрим, что же это за критерии.

00:02:26.000 --> 00:02:34.000

Самым первым делом, оно смотрит на количество инстансов в каждой availability зоне.

00:02:34.000 --> 00:02:53.000

Если в какой-то из availability зон больше всего инстансов, то она идет туда и смотрит в каком из инстансов в конкретной availability зоне самый старый launch configuration.

00:02:53.000 --> 00:02:59.000

Если такой находится, то она отключается.

00:02:59.000 --> 00:03:22.000

Если же у нас несколько availability зон с одинаковым количеством инстансов и находится несколько инстансов со старым launch configuration, то в этом случае она начинает сверять, какой из инстансов ближе к завершению billing hour.

00:03:22.000 --> 00:03:33.000

То есть оплата за EC2 инстанс происходит почасово, она округляется до часа, если меньше часа вы использовали EC2 инстанс.

00:03:33.000 --> 00:03:54.000

Поэтому для того, чтобы максимально использовать EC2 инстансы и сократить ваши расходы, она старается определить тот инстанс, который большую часть часа работала и с округлением вы потеряете минимальное количество минут в рамках этого часа.

00:03:54.000 --> 00:04:06.000

Если из двух или нескольких инстансов находится та, которая ближе всего, то она отключается.

00:04:06.000 --> 00:04:18.000

Если же таких инстансов по следующим критериям находится несколько, то Auto Scaling group останавливает инстанс в случайном порядке.

00:04:18.000 --> 00:04:30.000

Также вы можете видеть логику работы default termination policy Auto Scaling group на этой диаграмме.

00:04:30.000 --> 00:04:38.000

На этом мы разобрали вопрос Practice Test-а. Спасибо за внимание.