

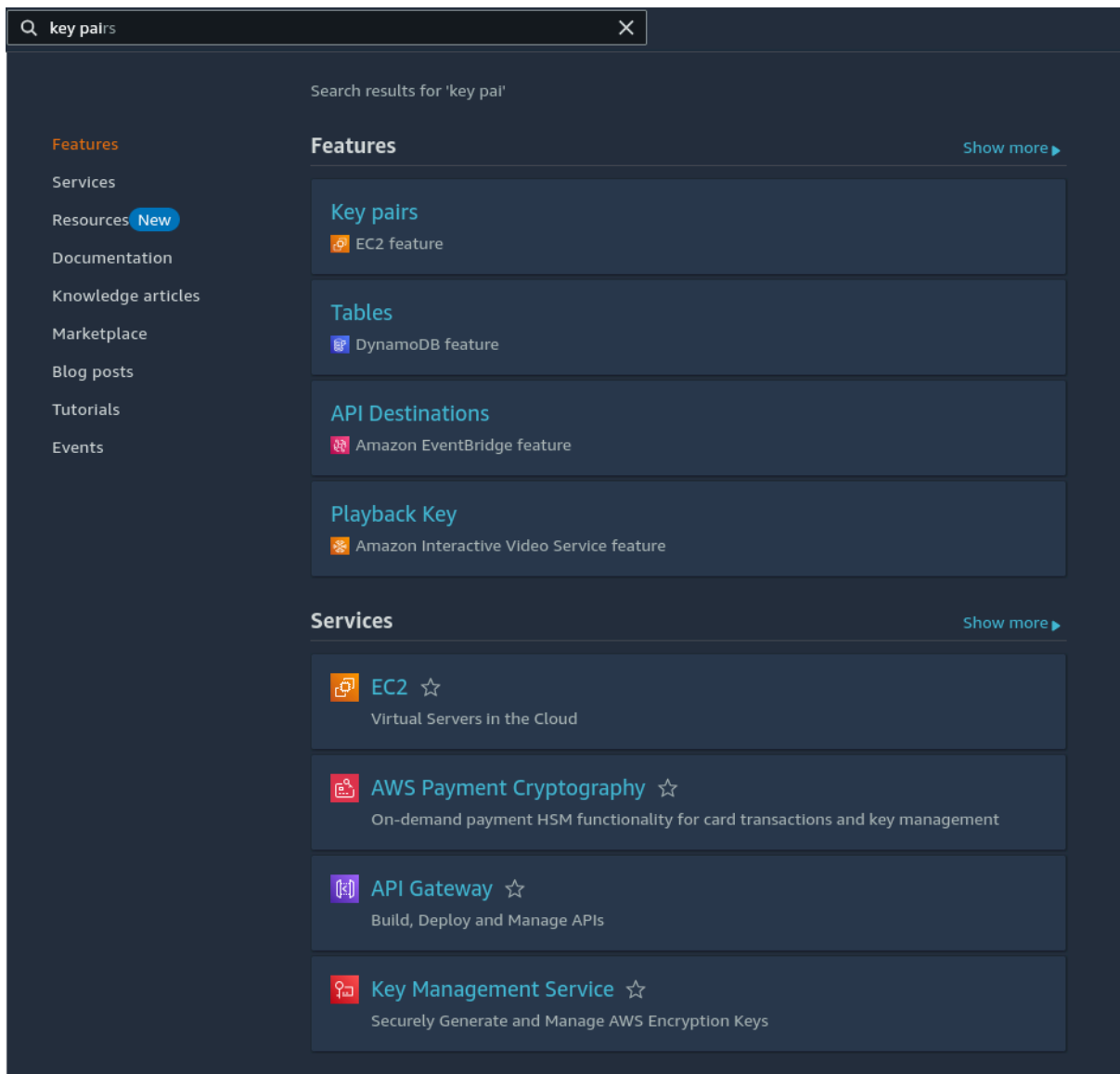
AMAZON EMR

Resumo

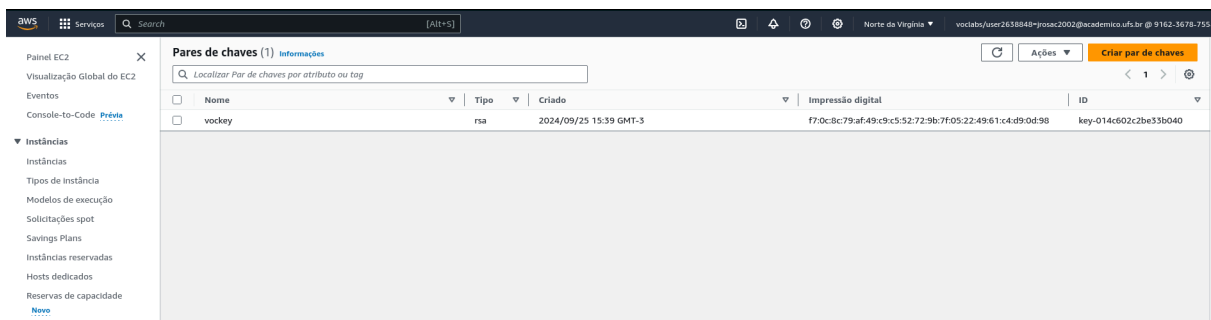
O presente projeto terá como objetivo demonstrar, de maneira detalhada, o processo de criação de um cluster utilizando o serviço Amazon EMR (Elastic MapReduce), explorando as etapas necessárias para configurar e iniciar o cluster. Em seguida, será abordada a execução de tarefas dentro desse ambiente. Por fim, será realizada a execução de um script Apache Spark no cluster. Este projeto busca fornecer uma visão prática e estruturada sobre o uso do Amazon EMR e o seu potencial no contexto de big data.

Primeira Etapa

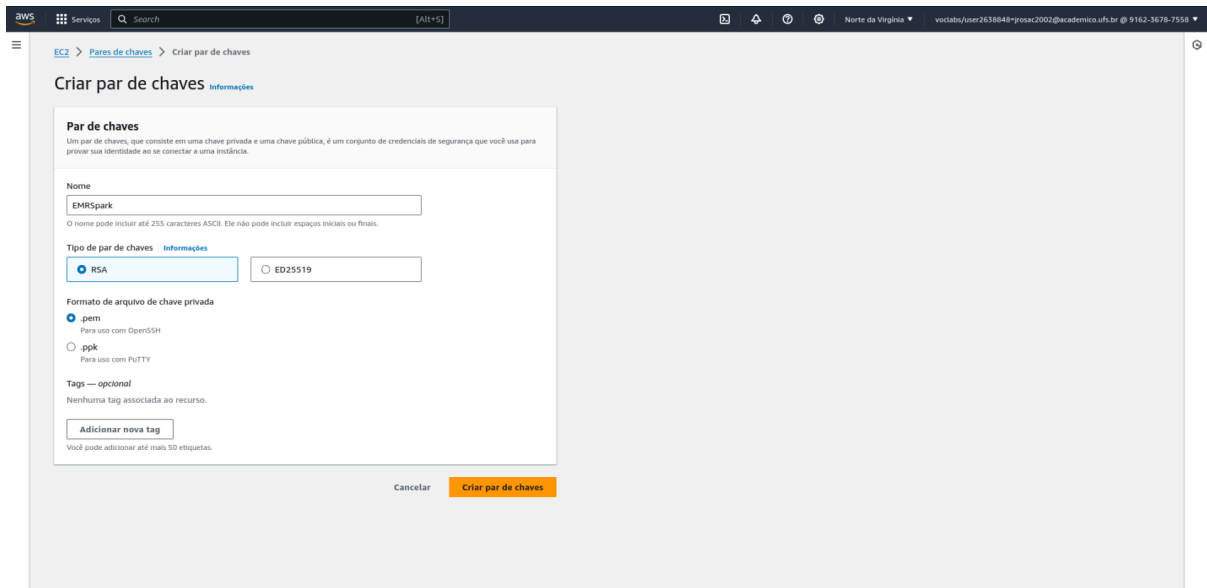
A primeira parte consiste em buscar e instanciar uma feature do EC2, chamada key pairs e com isso criar um par de chaves. O par de chaves é necessário pois ele provém ao Master Node a capacidade de conexão entre ele e os Workers(outros nodes). Sem esse par de chaves o Master Node não conseguirá executar as ações necessárias. Esse par de chaves será utilizado posteriormente para acessar o Cluster também.



Busca o do key pairs



Console para instanciar um par de chaves



Criar par de chaves [Informações](#)

Par de chaves
Um par de chaves, que consiste em uma chave privada e uma chave pública, é um conjunto de credenciais de segurança que você usa para provar sua identidade ao se conectar a uma instância.

Nome
EMRSpark
O nome pode incluir até 255 caracteres ASCII. Ele não pode incluir espaços iniciais ou finais.

Tipo de par de chaves [Informações](#)
☒ RSA ☐ ED25519

Formato de arquivo de chave privada
☒ .pem
Para uso com OpenSSH
☐ .ppk
Para uso com PuTTY

Tags — opcional
Nenhuma tag associada ao recurso.
[Adicionar nova tag](#)
Você pode adicionar até mais 50 etiquetas.

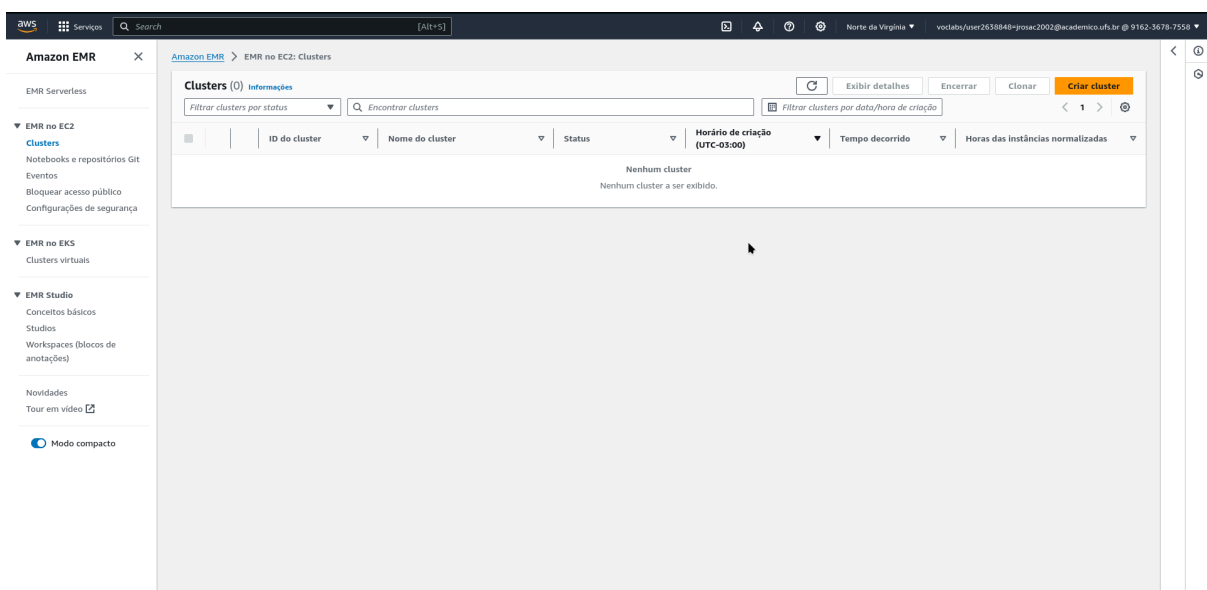
[Cancelar](#) [Criar par de chaves](#)

Informações para a criação do par de chaves

Após a criação do par de chaves será feito o download e arquivo .pem com o nome escolhido do par será salvo na sua máquina.

Segunda Etapa

A segunda etapa consiste na criação do cluster a partir do AMAZON EMR. A criação do cluster será dividida em 3 etapas principais: nome e aplicativos, Configuração do Cluster, provisionamento e escalabilidade.



Amazon EMR [Amazon EMR](#) > EMR no EC2: Clusters

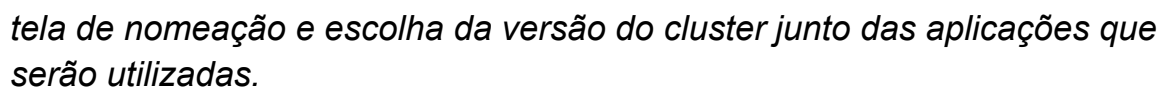
Clusters (0) [Informações](#)

[Filtrar clusters por status](#) [Filtrar clusters por data/hora de criação](#)

	ID do cluster	Nome do cluster	Status	Horário de criação (UTC-03:00)	Tempo decorrido	Horas das instâncias normalizadas
Nenhum cluster						
Nenhum cluster a ser exibido.						

[Exibir detalhes](#) [Encerrar](#) [Clonar](#) [Criar cluster](#)

Tela para a criação e visualização dos clusters criados.



tela de nomeação e escolha da versão do cluster junto das aplicações que serão utilizadas.

aws

Serviços

Search

[Alt+S]

Escolher ou inserir um ID de AMI

☒ Atualizar todos os pacotes instalados na reinicialização

▼ Configuração do cluster - obrigatório

Informações

Escolha um método de configuração para os grupos principal, núcleo e de nó de tarefa para o cluster.

☒ Grupos de instâncias uniformes

Escolha o mesmo tipo de instância do EC2 e a mesma opção de compra (Sob demanda ou Spot) para todos os nós do seu grupo de nós. Saiba mais

☐ Frotas de instâncias flexíveis

Escolha entre a maior variedade de opções de provisionamento para as instâncias do EC2 no seu cluster. Diversifique os tipos de instâncias e as opções de compra e use uma estratégia de alocação. Saiba mais

Grupos de instâncias uniformes

Primário

Escolher tipo de instância do EC2

m5.xlarge

4 vCore 16 GiB memória

Somente EBS armazenamento

Preço sob demanda: -

Preço spot mais baixo: -

Ações ▼

☐ Usar alta disponibilidade

Inicie clusters altamente disponíveis e mais resilientes com três nós primários em instâncias sob demanda. Essa configuração se aplica durante a vida útil do seu cluster. Saiba mais

► Configuração de nó - opcional

Núcleo

Escolher tipo de instância do EC2

m5.xlarge

4 vCore 16 GiB memória

Somente EBS armazenamento

Preço sob demanda: -

Preço spot mais baixo: -

Ações ▼

► Configuração de nó - opcional

Tarefa 1 de 2

Nome

Tarefa - 1

Escolher tipo de instância do EC2

m5.xlarge

4 vCore 16 GiB memória

Somente EBS armazenamento

Preço sob demanda: -

Preço spot mais baixo: -

Ações ▼

Remover grupo de instâncias

Resumo

Informações

Nome e aplicativos - obrigatório

Nome

EMRSpark

Versão do Amazon EMR

emr-6.3.0

Pacote de aplicativos

Spark (Spark 3.1.1, Zeppelin 0.9.0)

Configuração do cluster - obrigatório

Grupos de instâncias uniformes

Primário (m5.xlarge), Núcleo (m5.xlarge), Tarefa (m5.xlarge)

Provisionamento e escalabilidade do cluster - obrigatório

Configuração de provisionamento

Tamanho do núcleo: 1 instância

Tamanho da tarefa: 2 instâncias

Configurar perfil de serviço

Você deve escolher um perfil de serviço antes de criar esse cluster.

Escolher um perfil de serviço

Cancelar

Criar cluster

CloudShell

Comentários

Tela para a configuração do Cluster, definição do nome e capacidade de memória do master node(núcleo) e dps workers(tarefa e serviço).

Escalabilidade do Cluster, opções de rede e algumas configurações opcionais como Etapas.

Escalabilidade do Cluster, opções de rede e algumas configurações opcionais como Etapas.

Grupos de segurança do EC2 (firewall)

Etapas (0)

Informações

Remover

Editar

Adicionar

Use comandos e scripts para informar ao cluster onde encontrar e como processar seus dados. As etapas são executadas consecutivamente, a menos que você ative a opção Concorrência.

Encerramento do cluster e substituição dos nós

Informações

Escolha as configurações de encerramento e proteja seu cluster contra um desligamento acidental.

Opção de encerramento

☐ Encerrar o cluster manualmente

☐ Encerrar automaticamente o cluster após o término da última etapa

☒ Encerrar automaticamente o cluster após o tempo de inatividade (recomendado)

Tempo ocioso

Insira o tempo até o encerramento do cluster.

0 dia

01:00:00

Escolha um tempo maior que 1 minuto (00:01:00) e menor que 7 dias. O tempo está no formato hh:mm:ss (24 horas).

☐ Usar proteção contra encerramento

Protege seu cluster contra encerramento acidental. Se ativado, primeiro você deve desativar a proteção para encerrar o cluster. Recomendamos ativar a proteção contra encerramento para seus clusters de longa execução.

Substituição insalubre de nós - novo

Informações

☒ Ativar

O Amazon EMR interrompe normalmente os processos em nós não íntegros para minimizar a perda de dados e as interrupções de trabalho. Ele substitui rapidamente nós não íntegros por novas instâncias do EC2 para manter seus trabalhos funcionando sem problemas.

☐ Desativar

O Amazon EMR adiciona nós não íntegros a uma lista de negações enquanto os mantém no cluster, permitindo que você tenha acesso contínuo para solucionar problemas.

Ações de bootstrap (0)

Informações

Remover

Editar

Adicionar

Use ações de bootstrap para instalar software ou personalizar a configuração da instância.

Logs de cluster

Informações

Escolha onde e como armazenar seus arquivos de log.

Tags

Informações

Use etiquetas para pesquisar e filtrar recursos, e rastreie os custos da AWS associados ao seu cluster.

Configurações de software

Informações

Substitua as configurações padrão para aplicativos específicos em seu cluster.

Resumo

Informações

Nome e aplicativos - obrigatório

Nome

EMRSpark

Versão do Amazon EMR

emr-6.3.0

Pacote de aplicativos

Spark (Spark 3.1.1, Zeppelin 0.9.0)

Configuração do cluster - obrigatório

Grupos de instâncias uniformes

Primário (m5.xlarge), Núcleo (m5.xlarge), Tarefa (m5.xlarge)

Provisionamento e escalabilidade do cluster - obrigatório

Configuração de provisionamento

Tamanho do núcleo: 1 instância

Tamanho da tarefa: 2 instâncias

Configurar perfil de serviço

Você deve escolher um perfil de serviço antes de criar esse cluster.

Escolher um perfil de serviço

Cancelar

Criar cluster

CloudShell

Comentários

Opção de encerramento automático do cluster após 1 hora de uso, pode ser editado para o tempo desejado.

aws

Serviços

Search

[Alt+S]

Insira o tempo até o encerramento do cluster.

0 dia

01:00:00

Escolha um tempo maior que 1 minuto (00:01:00) e menor que 7 dias. O tempo está no formato hh:mm:ss (24 horas).

☐

Usar proteção contra encerramento

Protege seu cluster contra encerramento acidental. Se ativado, primeiro você deve desativar a proteção para encerrar o cluster. Recomendamos ativar a proteção contra encerramento para seus clusters de longa execução.

Substituição insalubre de nós - novo

Informações

☒

Ativar

O Amazon EMR interrompe normalmente os processos em nós não íntegros para minimizar a perda de dados e as interrupções de trabalho. Ele substitui rapidamente nós não íntegros por novas instâncias do EC2 para manter seus trabalhos funcionando sem problemas.

☐

Desativar

O Amazon EMR adiciona nós não íntegros a uma lista de negações enquanto os mantém no cluster, permitindo que você tenha acesso contínuo para solucionar problemas.

Ações de bootstrap (0)

Informações

Remover

Editar

Adicionar

Use ações de bootstrap para instalar software ou personalizar a configuração da instância.

Logs de cluster

Informações

Escolha onde e como armazenar seus arquivos de log.

Arquivamos automaticamente seus arquivos de log no Amazon S3. Você pode especificar sua própria localização do S3 ou usar a localização padrão do S3 para o Amazon EMR. O local padrão do registro é pré-preenchido no Localização do Amazon S3 campo.

☒

Publicar logs específicos do cluster no Amazon S3

Localização do Amazon S3

Q

s3://aws-logs-916236787558-us-east-1/elasticmapreduce

X

Exibir

Navegar no S3

Formato: usar s3://bucket/prefixo

☐

Criptografar logs específicos do cluster

Tags

Informações

Use etiquetas para pesquisar e filtrar recursos, e rastreie os custos da AWS associados ao seu cluster.

Configurações de software

Informações

Substitua as configurações padrão para aplicativos específicos em seu cluster.

Configuração de segurança e par de chaves do EC2

Informações

Escolha uma configuração de segurança ou crie uma nova para você reutilizar com outros clusters.

Configuração de segurança

Selecione as configurações da criptografia, autenticação, autorização e serviço de metadados da instância do cluster.

Resumo

Informações

Nome e aplicativos - obrigatório

Nome

EMRSpark

Versão do Amazon EMR

emr-6.3.0

Pacote de aplicativos

Spark (Spark 3.1.1, Zeppelin 0.9.0)

Configuração do cluster - obrigatório

Grupos de instâncias uniformes

Primário (m5.xlarge), Núcleo (m5.xlarge), Tarefa (m5.xlarge)

Provisionamento e escalabilidade do cluster - obrigatório

Configuração de provisionamento

Tamanho do núcleo: 1 instância

Tamanho da tarefa: 2 instâncias

Configurar perfil de serviço

Você deve escolher um perfil de serviço antes de criar esse cluster.

Escolher um perfil de serviço

Cancelar

Criar cluster

Logs do cluster salvos numa instância do AMAZON EC3, pode ser uma instância já existente, se não será instanciada uma.

aws

Serviços

Search

[Alt+S]

▼ Configuração de segurança e par de chaves do EC2

Informações

Escolha uma configuração de segurança ou crie uma nova para você reutilizar com outros clusters.

Configuração de segurança

Selecione as configurações da criptografia, autenticação, autorização e serviço de metadados da instância do cluster.

Escolher uma configuração de segurança

↺

Navegar

Criar configuração de segurança

Par de chaves do Amazon EC2 para o SSH do cluster

Informações

EMRSpark

×

Navegar

Criar par de chaves

▼ Perfis do Identity and Access Management (IAM) - obrigatório

Informações

Escolha ou crie um perfil de serviço e um perfil de instância para as instâncias do EC2 no cluster.

Perfil de serviço do Amazon EMR

Informações

O perfil de serviço é um perfil do IAM que o Amazon EMR assume para provisionar recursos e executar ações de nível de serviço com outros serviços da AWS.

Escolha um perfil de serviço existente

Selecione um perfil de serviço padrão ou um perfil personalizado com políticas do IAM anexadas para que o cluster possa interagir com outros serviços da AWS.

Escolha um perfil de serviço

Deixe que o Amazon EMR crie um novo perfil de serviço para que você possa conceder e restringir o acesso a recursos em outros serviços da AWS.

Função de serviço

EMR_AutoScaling_DefaultRole

↺

Perfil de instância do EC2 para o Amazon EMR

O perfil de instância atribui um perfil a cada instância do EC2 em um cluster. O perfil de instância deve especificar um perfil que possa acessar os recursos para as etapas e ações de bootstrap.

Escolha um perfil de instância existente

Selecione um perfil padrão ou um perfil de instância personalizado com políticas do IAM anexadas para que o cluster possa interagir com seus recursos no Amazon S3.

Escolha um perfil de instância

Deixe que o Amazon EMR crie um novo perfil de instância para que você possa especificar um conjunto personalizado de recursos para acesso no Amazon S3.

Perfil de instância

EMR_EC2_DefaultRole

↺

Função personalizada de ajuste de escala automático - opcional

Quando uma regra personalizada de ajuste de escala automático é acionada, o Amazon EMR assume essa função para adicionar e encerrar instâncias do EC2. Saiba mais

Função personalizada de ajuste de escala automático

EMR_AutoScaling_DefaultRole

↺

Criar perfil do IAM

Resumo

Informações

Nome e aplicativos - obrigatório

Nome

EMRSpark

Versão do Amazon EMR

emr-6.3.0

Pacote de aplicativos

Spark (Spark 3.1.1, Zeppelin 0.9.0)

Configuração do cluster - obrigatório

Grupos de instâncias uniformes

Primário (m5.xlarge), Núcleo (m5.xlarge), Tarefa (m5.xlarge)

Provisionamento e escalabilidade do cluster - obrigatório

Configuração de provisionamento

Tamanho do núcleo: 1 instância

Tamanho da tarefa: 2 instâncias

Cancelar

Criar cluster

CloudShell

Comentários

Perfis de serviço e instâncias como default para usar o próprio autoscaling da AWS.

Resumo das informações e configurações do Cluster criado

Resumo			
Informações do cluster		Aplicativos	Gerenciamento de clusters
ID do cluster j-NDATDZVFHTPG		Versão do Amazon EMR emr-6.3.0	Destino do log no Amazon S3 aws-logs-916236787558-us-east-1/elasticmapreduce
Configuração do cluster Grupos de instâncias		Aplicações instaladas Spark 3.1.1, Zeppelin 0.9.0	DNS público do nó primário -
Capacidade 1 Primário 1 Núcleo 2 Tarefa			Status e hora
			Status Iniciando
			Horário de criação 25 de setembro de 2024 às 16:36 (UTC-03:00)
			Tempo decorrido 0 segundos
Propriedades Ações de bootstrap Instâncias (hardware) Etapas Aplicativos Configurações Monitoramento Eventos Tags (0)			
Logs de cluster Informações		Encerramento do cluster e substituição dos nós Informações Editar	
Arquivar arquivos de log no Amazon S3 Ativado		Opção de encerramento Encerrar automaticamente o cluster após o tempo de inatividade	
Localização do Amazon S3 s3://aws-logs-916236787558-us-east-1/elasticmapreduce/		Tempo ocioso 1 hora	
Criptografia para logs Desativado		Proteção contra encerramento Desativado	
		Substituição insalubre de nós Ativado	
Rede e segurança Informações		Permissões	
Rede		Perfil de serviço para o Amazon EMR EMR_AutoScaling_DefaultRole	
Nuvem privada virtual (VPC) vpc-0fc7d8905d6167827		Perfil de instância do EC2 EMR_EC2_DefaultRole	
Sub-rede(s) e zona(s) de disponibilidade (AZ) subnet-004bbb0c26e0bf301 us-east-1c		Função personalizada de ajuste de escala automático EMR_AutoScaling_DefaultRole	
► Grupos de segurança do EC2 (firewall)			
Configuração de segurança			
Configuração de segurança Nenhuma			
Par de chaves do EC2 EMRSpark			

Terceira Etapa

A terceira etapa consiste em adicionar uma etapa dentro do EMR a partir de um script armazenado no EC3.

aws

Serviços

Q Search

[Alt+S]

Amazon S3

Buckets

Criar bucket

Criar bucket

Informações

Buckets são contêineres para dados armazenados no S3.

Configuração geral

Região da AWS

Leste dos EUA (Norte da Virgínia) us-east-1

Tipo de bucket

Informações

Propósito geral

Recomendados para a maioria dos casos de uso e padrões de acesso. Os buckets de uso geral são do tipo original do S3. Eles permitem uma combinação de classes de armazenamento que armazenam objetos de maneira redundante em várias zonas de disponibilidade.

Diretório

Recomendados para casos de uso de baixa latência. Esses buckets usam somente a classe de armazenamento do S3 Express One Zone, que fornece processamento mais rápido de dados em uma única zona de disponibilidade.

Nome do bucket

Informações

sparkaws

O nome do bucket deve ser exclusivo no namespace global e seguir as regras de nomenclatura do bucket. [Veja as regras para nomenclatura de buckets](#)

Copiar configurações do bucket existente - *opcional*

Somente as configurações de bucket na configuração a seguir são copiadas.

Escolher bucket

Formato: s3://bucket/prefix

Propriedade de objeto

Informações

Controle a propriedade de objetos gravados nesse bucket a partir de outras contas da AWS e o uso de listas de controle de acesso (ACLs). A propriedade do objeto determina quem pode especificar o acesso aos objetos.

ACLs desabilitadas (recomendado)

Todos os objetos nesse bucket são de propriedade dessa conta. O acesso a esse bucket e seus objetos é especificado usando apenas políticas.

ACLs habilitadas

Os objetos nesse bucket podem ser de propriedade de outras contas da AWS. O acesso a esse bucket e seus objetos pode ser especificado usando ACLs.

Propriedade do objeto

Imposto pelo proprietário do bucket

Configurações de bloqueio do acesso público deste bucket

O acesso público é concedido a buckets e objetos por meio de listas de controle de acesso (ACLs), políticas de bucket, políticas de ponto de acesso ou todas elas. Para garantir que o acesso público a este bucket e todos os seus objetos seja bloqueado, ative a opção de Bloquear todo o acesso público. Essas configurações serão aplicadas apenas a este bucket e aos respectivos pontos de acesso. A AWS recomenda ativar a opção Bloquear todo o acesso público. Porém, antes de aplicar qualquer uma dessas configurações, verifique se as aplicações funcionarão corretamente sem acesso público. Caso precise de algum nível de acesso público a este bucket ou aos objetos que ele contém, é possível personalizar as configurações individuais abaixo para que atendam aos seus casos de uso de armazenamento específicos. [Saiba](#)

CloudShell

Comentários

aws

Serviços

Search

[Alt+S]

Versionamento de bucket

O versionamento é um meio de manter múltiplas variantes de um objeto no mesmo bucket. Você pode usar o versionamento para preservar, recuperar e restaurar todas as versões de cada objeto armazenado no bucket do Amazon S3. Com o versionamento, você pode recuperar facilmente ações não intencionais do usuário e falhas da aplicação. [Saiba mais](#)

Versionamento de bucket

☒ Desativar

☐ Ativar

Tags - opcional (0)

Você pode usar tags de bucket para rastrear custos de armazenamento e organizar buckets. [Saiba mais](#)

Nenhuma tag associada a este bucket.

Adicionar tag

Criptografia padrão [Informações](#)

A criptografia no lado do servidor é aplicada automaticamente a novos objetos armazenados nesse bucket.

Tipo de criptografia [Informações](#)

☒ Criptografia do lado do servidor com chaves gerenciadas do Amazon S3 (SSE-S3)

☐ Criptografia do lado do servidor com chaves do AWS Key Management Service (SSE-KMS)

☐ Criptografia de duas camadas no lado do servidor com chaves do AWS Key Management Service (DSSE-KMS)

Proteja seus objetos com duas camadas separadas de criptografia. Para obter detalhes sobre a precificação, consulte os [preços do DSSE-KMS](#) na guia Armazenamento da [página de preços do Amazon S3](#).

Chave do bucket

O uso de uma chave de bucket do S3 para SSE-KMS reduz os custos de criptografia ao diminuir as chamadas para o AWS KMS. As chaves de bucket do S3 não são compatíveis com o DSSE-KMS. [Saiba mais](#)

☐ Desativar

☒ Ativar

► Configurações avançadas

Depois de criar o bucket, você pode fazer upload de arquivos e pastas para o bucket e definir configurações adicionais do bucket.



Cancelar


Criar bucket

CloudShell


Comentários

Primeiramente será necessário instanciar um bucket no Amazon S3, nele ficará armazenado o script a ser processado no cluster. As configurações utilizadas serão as recomendadas e já marcadas por default.

  Serviços

 Search

[Alt+S]



[Amazon S3](#) > [Buckets](#) > [sparkaws](#) > Carregar

Carregar [Informações](#)

Adicione os arquivos e pastas que você deseja carregar no S3. Para fazer upload de um arquivo maior que 160 GB, use a AWS CLI, o SDK da AWS ou a API REST do Amazon S3. [Saiba mais](#)

Arraste e solte aqui os arquivos e pastas para upload ou selecione **Adicionar arquivos** ou **Adicionar pastas**.


Arquivos e pastas (0)

Remover

Adicionar arquivos

Adicionar pasta

Todos os arquivos e pastas desta tabela serão carregados.

 Encontrar por nome

< 1 >

☐

Nome

▼

Pasta

Nenhum arquivo ou pasta

Você não escolheu qualquer arquivo ou pasta para carregar.

Destino [Informações](#)

Destino

[s3://sparkaws](#)

► **Detalhes do destino**

Configurações de bucket que afetam novos objetos armazenados no destino especificado.

► **Permissões**

Conceda acesso público e acesso a outras contas da AWS.

► **Propriedades**

Especifique classe de armazenamento, configurações de criptografia, tags e muito mais.

Cancelar

Carregar

 CloudShell

Comentários

Após isso você irá selecionar o arquivo da sua máquina que deseja armazenar no S3, e fazer o upload para o bucket instanciado, no caso da atividade estou utilizando um arquivo chamado SparkNaAws.py.

Amazon S3 > Buckets > sparkaws > SparkNaAws.py

SparkNaAws.py

Informações

Copiar URI do S3 Fazer download Abrir Ações de objeto

Propriedades Permissões Versões

Visão geral do objeto

Proprietário awslabs:ow25113871625144762	URI do S3 s3://sparkaws/SparkNaAws.py
Região da AWS Leste dos EUA (Norte da Virgínia) us-east-1	Nome de recurso da Amazon (ARN) arn:aws:s3::sparkaws/SparkNaAws.py
Última modificação 26 Sep 2024 01:01:53 AM -03	Tag da entidade (Etag) 19df9e10e35944c8b457535b84ec9385
Tamanho 518,0 B	URL de objeto https://sparkaws.s3.amazonaws.com/SparkNaAws.py
Tipo py	
Chave SparkNaAws.py	

Visão geral do gerenciamento de objetos

As propriedades do bucket e as configurações de gerenciamento de objetos a seguir afetam o comportamento desse objeto.

Propriedades de bucket

Versionamento de bucket

Quando habilitado, várias variantes de um objeto podem ser armazenadas no bucket visando facilitar a recuperação de ações não intencionais do usuário e falhas da aplicação.

Desabilitado

O versionamento do bucket "sparkaws" não está ativado

Recomendamos que você habilite o versionamento de bucket para ajudar a proteger contra substituição ou exclusão involuntária de objetos. [Saiba mais](#)

Habilitar versionamento de bucket

Configurações de gerenciamento

Status da replicação

Quando uma regra de replicação é aplicada a um objeto, o status de replicação indica o andamento da operação.

Exibir regras de replicação

Regra de expiração

Você pode usar uma configuração de ciclo de vida para definir regras de expiração para programar a remoção desse objeto após um período predefinido.

Data de expiração

O objeto será excluído permanentemente nessa data.

Após o upload ser concluído, você poderá ver as informações do script dentro do bucket, após isso você irá copiar a URL como S3, apertando o botão no canto superior direito.

AWS Services [Alt+S]

Amazon EMR > EMR no EC2 Clusters > EMRAtv

EMRAtv

Atualizado há menos de um minuto Encerrar Clonar na AWS CLI Clonar

Resumo

Propriedades Ações de bootstrap Instâncias (hardware) Etapas Aplicativos Configurações Monitoramento Eventos Tags (0)

Etapas (3)

Informações

Cada etapa é uma unidade de trabalho que contém instruções para manipular dados processados pelo software instalado no cluster.

Etapas simultâneas: 1

Filtrar etapas por status Encerrar etapas

ID da etapa	Status	Nome	Arquivos de log	Horário de criação (UTC-03:00)	Horário de início (UTC-03:00)	Tempo decorrido
s-05279011GQXBKOZ559BN	Completed	exemplo etapa	controller syslog stderr stdout	26 de setembro de 2024 às 01:46	26 de setembro de 2024 às 01:47	28 segundos
s-0116559XBVYAHVSKGG	Completed	exemplo1	controller syslog stderr stdout	26 de setembro de 2024 às 01:41	26 de setembro de 2024 às 01:41	28 segundos
s-060133224M654MRED6ZW	Completed	Aplicativo do spark	controller syslog stderr stdout	26 de setembro de 2024 às 01:06	26 de setembro de 2024 às 01:07	1 minuto, 2 segundos

Após o upload no arquivo no S3, você irá retornar ao EMR, irá acessar o cluster instanciado, selecionará a opção etapa e irá apertar o botão de adicionar no canto superior direito

aws

Serviços

Search

[Alt+S]

Amazon EMR > EMR no EC2: Clusters > EMRAtv > Adicionar etapa

Adicionar etapa

Informações

Configurações da etapa

Tipo

☐ JAR personalizado

Adiciona uma etapa que permite escrever um script personalizado para processar seus dados usando a linguagem de programação Java.

☐ Programa de transmissão

Adiciona uma etapa que usa entrada padrão para executar scripts de mapeador/redutor e enviar resultados para a saída padrão.

☐ Programa do Hive

Adiciona uma etapa que envia um script do Hive para interações de data warehouse.

☒ Aplicativo do Spark

Adiciona uma etapa que envia trabalhos para a estrutura do Spark no cluster.

☐ Script de shell

Solucione problemas no seu cluster.

Nome

Inserir um nome

Modo de implantação

☐ Modo de cluster

Execute o driver em um nó de operador.

☒ Modo de cliente

Execute o driver no nó primário como cliente externo.

Local do aplicativo

Caminho para um arquivo JAR com o aplicativo e as dependências (modo de implantação de cliente é compatível apenas com um caminho local).

s3://exemploemr-1/SparkNaAws.py

Opções de Spark-submit - opcional

Especifique outras opções para spark-submit.

--class, org.apache.spark.examples.SparkPi

Argumentos - opcional

Informações

Especifique argumentos opcionais para o script.

/usr/lib/spark/examples/jars/spark-examples.jar10

Ação da etapa

Ação se a etapa falhar

A ação a ser executada quando a etapa falhar.

☒ Continuar

Prossegue para a próxima etapa na fila.

☐ Cancelar e aguardar

Cancela todas as etapas pendentes e retorna o cluster para o estado de espera.

☐ Encerrar cluster

Desliga o cluster.

Cancelar





Adicionar etapa

CloudShell

Comentários

Após isso selecione a opção aplicativo do spark e modo de implantação modo

cliente, para rodar o script no nó primário como cliente externo e em seguida aperte adicionar etapa

<input type="checkbox"/>	ID da etapa ▾	Status ▾	Nome ▾	Arquivos de log 	Horário de criação (UTC-03:00) ▾	Horário de Início (UTC-03:00)
<input type="checkbox"/>	 s-05279011GQXBKO25S98N	 Completed	exemplo etapa	controller syslog stderr stdout 	26 de setembro de 2024 às 01:46	26 de setembro de 2024 às 01:47

Após a etapa ser processada pelo cluster, ela irá retornar os logs de controller de stderr, que irão conter dados referentes ao processamento e possíveis erros e também o stdout, o qual irá conter a saída do script desejado.

Com isso você terá conseguido rodar um script de aplicação spark dentro da aws