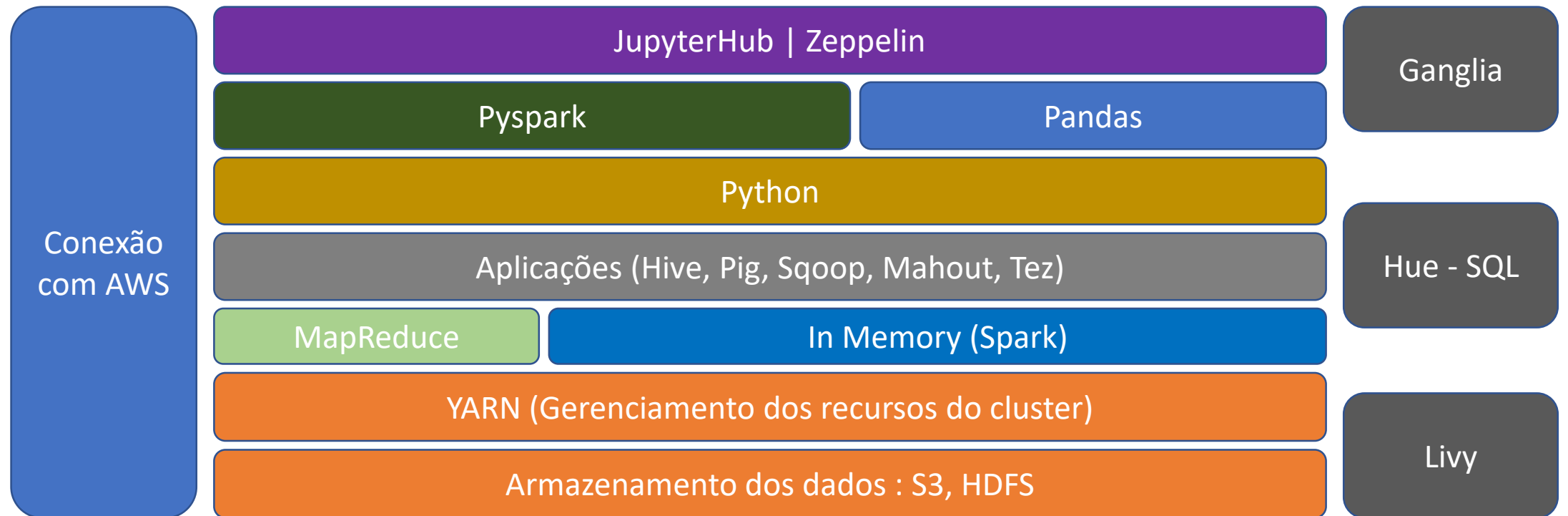
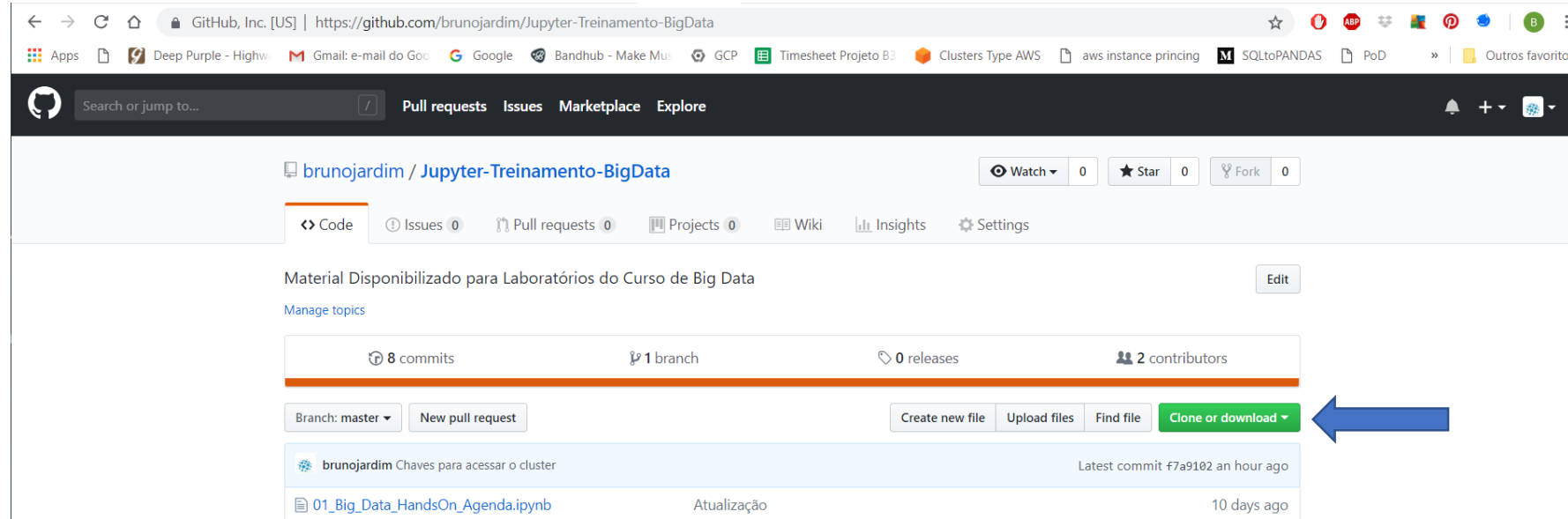


Arquitetura/Ecosystema Hadoop que vamos utilizar neste treinamento



Arquivos necessários para treinamento

1) Clonar repositório para pasta local: acessar Git e copiar URL do repositório



2) Acessar Git Bash (fazer download caso não tenha: <https://git-scm.com/downloads>)

3) Acesse uma pasta local do seu computador via Git Bash:

`Cd <caminho/pasta/local>`

`Git clone <URL copiada do Git(site)>`

`treinamento_2018@bigdata`

aws

Services

Resource Groups

IAM

EC2

EMR

Lambda

S3

RDS

CloudWatch

Dynan

bruno @ 9834-4566-1575

N. Virginia

Support

Amazon EMR

Clusters

Configuração de segurança

Sub-redes da VPC

Eventos

Ajuda

What's new

É possível usar o Catálogo de dados do AWS Glue como o metastore externo do Hive para cargas de trabalho do [Apache Spark](#), do [Apache Hive](#) e do [Presto](#) no Amazon EMR versão 5.10.0 e posterior. Para começar, basta selecionar o Catálogo de dados do AWS Glue para obter os metadados da tabela ao criar o cluster.

Criar cluster

Visualizar detalhes

Clonar

Encerrar

Filter: Clusters ativos

Filtrar clusters...

0 clusters (todos carregados)

	Nome	ID	Status	Horário de criação (UTC-2)	Tempo decorrido	Horas da instância normalizada
--	------	----	--------	----------------------------	-----------------	--------------------------------

[Services](#) ▾[Resource Groups](#) ▾[IAM](#)[EC2](#)[EMR](#)[Lambda](#)[S3](#)[RDS](#)[CloudWatch](#)[DynamoDB](#)

bruno @ 9834-4566-1575 ▾

N. Virginia ▾

[Support](#) ▾

Criar cluster - Opções rápidas

[Ir para opções avançadas](#)

Configuração geral

Nome do cluster

☒ Registro em log ⓘPasta do S3

Modo de execução



Cluster ⓘ



Execução da etapa ⓘ

Configuração de software

Versão



Aplicativos



Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4



HBase: HBase 1.4.7 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.3, Hue 4.2.0, Phoenix 4.14.0, and ZooKeeper 3.4.13



Presto: Presto 0.212 with Hadoop 2.8.5 HDFS and Hive 2.3.3 Metastore



Spark: Spark 2.3.2 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.0



Usar o catálogo de dados do AWS Glue para obter uma tabela de metadados ⓘ

[Services](#) ▾[Resource Groups](#) ▾[IAM](#)[EC2](#)[EMR](#)[Lambda](#)[S3](#)[RDS](#)[CloudWatch](#)[DynamoDB](#)[Alerts](#)

bruno @ 9834-4566-1575 ▾

N. Virginia ▾

[Support](#) ▾

Criar cluster - opções avançadas

[Ir para opções rápidas](#)

Step 1: Software e etapas

[Step 2: Hardware](#)[Step 3: Configurações gerais de cluster](#)[Step 4: Segurança](#)

Configuração de software

Versão ⓘ☒ Hadoop 2.8.5☒ JupyterHub 0.9.4☒ Ganglia 3.7.2☒ Hive 2.3.3☐ MXNet 1.3.0☒ Hue 4.2.0☒ Spark 2.3.2☒ Zeppelin 0.8.0☒ Tez 0.8.4☒ HBase 1.4.7☐ Presto 0.212☒ Sqoop 1.4.7☐ Phoenix 4.14.0☐ HCatalog 2.3.3☐ Livy 0.5.0☐ Flink 1.6.1☒ Pig 0.17.0☐ ZooKeeper 3.4.13☒ Mahout 0.13.0☒ Oozie 5.0.0☐ TensorFlow 1.11.0

Configurações do catálogo de dados do AWS Glue (opcional)

☐ Usar para metadados da tabela do Hive ⓘ☐ Usar para metadados da tabela do Spark ⓘ

Configurações do armazenamento de HBase ⓘ

Modo de armazenamento ☐ HDFS☒ S3Diretório raiz 📁☐ Usar como réplica de leitura ⓘ

Editar configurações de software ⓘ

☒ Inserir configuração ☐ Carregar JSON de S3



Services ▾

Resource Groups ▾

IAM

EC2

EMR

Lambda

S3

RDS

CloudWatch

DynamoDB



bruno @ 9834-4566-1575 ▾

N. Virginia ▾

Support ▾

Tamanho do volume do EBS do dispositivo raiz GiB ⓘ

Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. [Learn more about instance purchasing options](#)

Tipo de nó	Tipo de instância	Contagem de instâncias	Opção de compra	Auto Scaling
Principal Principal - 1 ⓘ	m4.xlarge ⓘ 8 vCore, 16 GiB de memória, armazenamento apenas EBS Armazenamento de EBS: 32 GiB ⓘ ⓘ	1 Instâncias	<input type="radio"/> Sob demanda ⓘ <input checked="" type="radio"/> Spot ⓘ Set max price per instance/hr ▾ \$ 0.100 ▴ ▾	Indisponível para o Principal ⓘ
Serviços Serviços - 2 ⓘ	m4.large ⓘ 4 vCore, 8 GiB de memória, armazenamento apenas EBS Armazenamento de EBS: 32 GiB ⓘ ⓘ	<input type="text" value="0"/> Instâncias	<input checked="" type="radio"/> Sob demanda ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▾	Não ativado ⓘ
Tarefa ✕ Tarefa - 3 ⓘ	m4.large ⓘ 4 vCore, 8 GiB de memória, armazenamento apenas EBS Armazenamento de EBS: 32 GiB ⓘ ⓘ	<input type="text" value="0"/> Instâncias	<input checked="" type="radio"/> Sob demanda ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▾	Não ativado ⓘ

[+ Adicionar grupo de instâncias de tarefas](#)

Cancel

Previous

Next

[Services](#) ▾[Resource Groups](#) ▾[IAM](#)[EC2](#)[EMR](#)[Lambda](#)[S3](#)[RDS](#)[CloudWatch](#)[DynamoDB](#)

bruno @ 9834-4566-1575 ▾

N. Virginia ▾

[Support](#) ▾

Criar cluster - opções avançadas

[Ir para opções rápidas](#)[Step 1: Software e etapas](#)[Step 2: Hardware](#)**Step 3: Configurações gerais de cluster**[Step 4: Segurança](#)

Opções gerais

Nome do cluster ☒ Registro em log ⓘPasta do S3 ☒ Depuração ⓘ☒ Proteção contra encerramento ⓘ

Tags ⓘ

Chave	Valor (opcional)	
<input type="text" value="Nome"/>	<input type="text" value="BrunoJardim"/>	✕
<input type="text" value="Treinamento"/>	<input type="text" value="BigData"/>	✕
<input type="text" value="Adicionar uma chave para criar uma tag"/>	<input type="text"/>	

Opções adicionais

☐ Visualização consistente do EMRFS ⓘID personalizado de AMI ⓘ

[Services](#) ▾[Resource Groups](#) ▾[IAM](#)[EC2](#)[EMR](#)[Lambda](#)[S3](#)[RDS](#)[CloudWatch](#)[DynamoDB](#)

bruno @ 9834-4566-1575 ▾

N. Virginia ▾

[Support](#) ▾

Criar cluster - opções avançadas

[Ir para opções rápidas](#)[Step 1: Software e etapas](#)[Step 2: Hardware](#)[Step 3: Configurações gerais de cluster](#)**Step 4: Segurança**

Opções de segurança

Par de chaves EC2 BigDataEMR ▾ ⓘ☒ Cluster visível a todos os usuários do IAM na conta ⓘ

Permissões ⓘ

☒ Padrão ☐ Personalizado

Use as funções padrão do IAM. Caso não haja funções, elas serão criadas automaticamente com políticas gerenciadas para atualizações automáticas de políticas.

Função do EMR [EMR_DefaultRole](#) ⓘPerfil de instância do EC2 [EMR_EC2_DefaultRole](#) ⓘFunção do Auto Scaling [EMR_AutoScaling_DefaultRole](#) ⓘ[▶ Autenticação e criptografia](#)[▶ Grupos de segurança do EC2](#)[Cancel](#)[Previous](#)[Criar cluster](#)



Services ▾

Resource Groups ▾

IAM

EC2

EMR

Lambda

S3

RDS

CloudWatch

Dynan



bruno @ 9834-4566-1575 ▾

N. Virginia ▾

Support ▾

Amazon EMR

Clusters

Configuração de
segurança

Sub-redes da VPC

Eventos

Ajuda

What's new

Clonar

Encerrar

Exportação de CLI da AWS

Cluster: EMR_BJ Iniciando

Resumo

Histórico do aplicativo

Monitoramento

Hardware

Eventos

Etapas

Configurações

Ações de bootstrap

Conexões: --

DNS público principal: --

Tags: Nome = BrunoJardim, Treinamento = BigData [Visualizar todas/Editar](#)

Resumo

ID: j-1Y3QBHYPWEM17

Data de criação: 2018-11-17 15:16 (UTC-2)

Tempo decorrido: 1 segundo

Encerramento Não
automático:

Proteção contra Ativado [Alterar](#)
encerramento:

Detalhes da configuração

Rótulo da versão: emr-5.19.0

Distribuição do Amazon 2.8.5

Hadoop:

Aplicativos: JupyterHub 0.9.4, Ganglia 3.7.2, Hive
2.3.3, Hue 4.2.0, Spark 2.3.2,
Zeppelin 0.8.0, Tez 0.8.4, Sqoop
1.4.7, Pig 0.17.0, Mahout 0.13.0,
Oozie 5.0.0, HBase 1.4.7

URI do log: s3://aws-logs-983445661575-us-
east-1/elasticmapreduce/

Visualização Desativado
consistente do

EMRFS:

ID personalizado de --

AMI:

Rede e hardware

Zona de --
disponibilidade:

ID da sub-rede: [subnet-3980235e](#)

Principal: Provisionamento 1 m4.xlarge
Spot (max \$0.10/hr)

Serviços: --

Tarefa: --

Segurança e acesso

Nome da chave: BigDataEMR

Perfil da instância do EMR_EC2_DefaultRole
EC2:

Amazon EMR

Clusters

Configuração de segurança

Sub-redes da VPC

Eventos

Ajuda

What's new

É possível usar o Catálogo de dados do AWS Glue como o metastore externo do Hive para cargas de trabalho do Apache Spark, do Apache Hive e do Presto no Amazon EMR versão 5.10.0 e posterior. Para começar, basta selecionar o Catálogo de dados do AWS Glue para obter os metadados da tabela ao criar o cluster.

Criar cluster

Visualizar detalhes

Clonar

Encerrar

Filter: Todos os clusters

Filtrar clusters carregados...

100 clusters carregados

load more

	Nome	ID	Status	Horário de criação (UTC-2)	Tempo decorrido	Horas da instância normalizada
<input type="checkbox"/>	EMR_BJ	j-1Y3QBHYPWEM17	Aguardando Cluster pronto	2018-11-17 15:16 (UTC-2)	16 minutos	0

Resumo

DNS público principal ec2-35-169-124-144.compute-1.amazonaws.com

Proteção contra encerramento: Ativado

Tags: Nome = BrunoJardim, Treinamento = BigData

Hardware

Principal: Running 1 m4.xlarge Spot (max \$0.10/hr)

Serviços: --

Tarefa: --

Etapas

Nome	Status	Horário de início (UTC-2)	Tempo decorrido
Configurar depuração do Hadoop	Concluído	2018-11-17 15:30 (UTC-2)	2 segundos

Ações de bootstrap

Nome
Nenhum ação de bootstrap disponível

Visualizar detalhes do cluster

Visualizar detalhes do monitoramento



Services ▾

Resource Groups ▾

IAM

EC2

EMR

Lambda

S3

RDS

CloudWatch

Dynan



bruno @ 9834-4566-1575 ▾

N. Virginia ▾

Support ▾

Amazon EMR

Clusters

Configuração de segurança

Sub-redes da VPC

Eventos

Ajuda

What's new

Clonar

Encerrar

Exportação de CLI da AWS

Cluster: EMR_BJ

Aguardando

Cluster ready after last step completed.

Resumo

Histórico do aplicativo

Monitoramento

Hardware

Eventos

Etapas

Configurações

Ações de bootstrap

Conexões:

[Habilitar conexão da web](#) – Hue, Zeppelin, Servidor de histórico do Spark, Ganglia, HBase, JupyterHub, Gerenciador de recursos ... (Visualizar tudo)

DNS público principal

ec2-35-169-124-144.compute-1.amazonaws.com [SSH](#)

Tags:

Nome = BrunoJardim, Treinamento = BigData [Visualizar todas/Editar](#)

Resumo

ID: j-1Y3QBHYPWEM17

Data de criação: 2018-11-17 15:16 (UTC-2)

Tempo decorrido: 18 minutos

Encerramento automático: Não

Proteção contra encerramento: Ativado [Alterar](#)

Detalhes da configuração

Rótulo da versão: emr-5.19.0

Distribuição do Amazon 2.8.5

Hadoop:

Aplicativos: JupyterHub 0.9.4, Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Spark 2.3.2, Zeppelin 0.8.0, Tez 0.8.4, Sqoop 1.4.7, Pig 0.17.0, Mahout 0.13.0, Oozie 5.0.0, HBase 1.4.7

URI do log: s3://aws-logs-983445661575-us-east-1/elasticmapreduce/

Visualização consistente do EMRFS: Desativado

EMRFS:

ID personalizado de --
AMI:

Rede e hardware

Zona de us-east-1b
disponibilidade:

ID da sub-rede: [subnet-3980235e](#)

Principal: Running 1 m4.xlarge
Spot (max \$0.10/hr)

Serviços: --

Tarefa: --

aws

Services

Resource

Amazon EMR

Clusters

Configuração de segurança

Sub-redes da VPC

Eventos

Ajuda

What's new

Cluster: E

Clonar

Resumo

Conexões:

DNS público

Tags:

Resumo

Data de

Tempo d

Ence

au

Proteç

encer

Habilitar conexão da web

Configurar conexão da web

O Hadoop, o Ganglia e outros aplicativos publicam interfaces de usuário como sites hospedados no nó principal. Por razões de segurança, esses sites estão disponíveis apenas no servidor web local do nó principal.

Para se conectar às interfaces da web, você deve estabelecer um túnel SSH com o nó principal, usando o encaminhamento de portas dinâmicas ou locais. Se você estabelecer um túnel SSH usando o encaminhamento de portas dinâmicas, também deverá configurar um servidor de proxy para visualizar as interfaces da web.

Etapa 1: Abrir um túnel SSH para o nó principal do Amazon EMR - [Saiba mais](#)

WindowsMac/Linux

1. Faça download de PuTTY.exe para seu computador de: <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>
2. Inicie o PuTTY.
3. Na lista Categoria, clique em Sessão
4. No campo Nome de host, digite **hadoop@ec2-35-169-124-144.compute-1.amazonaws.com**
5. Na lista Categoria, expanda Conexão > SSH > Autenticar
6. Para autenticação no arquivo de chave privada, clique em Procurar e selecione o arquivo de chave privada (**BigDataEMR.ppk**) usado para ativar o cluster.
7. Na lista Categoria, expanda Conexão > SSH e depois clique em Túneis.
8. No campo Porta de origem, digite **8157** (uma porta local não utilizada e escolhida aleatoriamente).
9. Selecione as opções Dinâmica e Automática.
10. Deixe o campo Destino vazio e clique em Adicionar.
11. Clique em Abrir.
12. Clique em Sim para descartar o alerta de segurança.

aws

Services

Resource

Amazon EMR

Clusters

Configuração de segurança

Sub-redes da VPC

Eventos

Ajuda

What's new

Cluster: E

Resumo

Conexões:

DNS público

Tags:

Resumo

Data de

Tempo de

Encer

Proteç

encer

Segurança

Habilitar conexão da web

Etapa 2: Configurar uma ferramenta de gerenciamento de proxy - Saiba mais

ChromeFirefox

1. Faça download da versão padrão de FoxyProxy e instale-a de:
<http://foxyproxy.mozdev.org/downloads.html>

2. Reinicie o Chrome depois de instalar o FoxyProxy.

3. Usando um editor de texto, crie um arquivo chamado foxyproxy-settings.xml contendo o seguinte:

```
<?xml version="1.0" encoding="UTF-8"?>
<foxyproxy>
  <proxies>
    <proxy name="emr-socks-proxy" id="2322596116" notes="" fromSubscription="false" enabled="true" mode="manual"
selectedTabIndex="2" lastresort="false" animatedIcons="true" includeInCycle="true" color="#0055E5" proxyDNS="true"
noInternalIPs="false" autoconfMode="pac" clearCacheBeforeUse="false" disableCache="false" clearCookiesBeforeUse="false"
rejectCookies="false">
      <matches>
        <match enabled="true" name="*ec2*.amazonaws.com*" pattern="*ec2*.amazonaws.com*" isRegex="false"
isBlackList="false" isMultiline="false" caseSensitive="false" fromSubscription="false" />
        <match enabled="true" name="*ec2*.compute*" pattern="*ec2*.compute*" isRegex="false" isBlackList="false"
isMultiline="false" caseSensitive="false" fromSubscription="false" />
        <match enabled="true" name="10.*" pattern="http://10.*" isRegex="false" isBlackList="false"
isMultiline="false" caseSensitive="false" fromSubscription="false" />
        <match enabled="true" name="*10*.amazonaws.com*" pattern="*10*.amazonaws.com*" isRegex="false"
isBlackList="false" isMultiline="false" caseSensitive="false" fromSubscription="false" />
        <match enabled="true" name="*10*.compute*" pattern="*10*.compute*" isRegex="false" isBlackList="false"
isMultiline="false" caseSensitive="false" fromSubscription="false" />
        <match enabled="true" name="*.compute.internal*" pattern="*.compute.internal*" isRegex="false"
isBlackList="false" isMultiline="false" caseSensitive="false" fromSubscription="false" />
        <match enabled="true" name="*.ec2.internal*" pattern="*.ec2.internal*" isRegex="false" isBlackList="false"
isMultiline="false" caseSensitive="false" fromSubscription="false" />
      </matches>
      <manualconf host="localhost" port="8157" socksVersion="5" isSocks="true" username="" password="" domain="" />
    </proxy>
  </proxies>
</foxyproxy>
```

4-4566-1575

N. Virginia

Support

aws

Services

Resource Groups

Amazon EMR

Clusters

Configuração de segurança

Sub-redes da VPC

Eventos

Ajuda

What's new

Cluster: E

Clonar

Resumo

Conexões:

DNS público

Tags:

Resumo

Data de

Tempo d

Ence

au

Proteç

encer

Segurança e

Nome

Habilitar conexão da web

```
isBlackList="false" isMultiline="false" caseSensitive="false" fromSubscription="false" />
    <match enabled="true" name="*ec2*.compute*" pattern="*ec2*.compute*" isRegex="false" isBlackList="false"
isMultiline="false" caseSensitive="false" fromSubscription="false" />
    <match enabled="true" name="10.*" pattern="http://10.*" isRegex="false" isBlackList="false"
isMultiline="false" caseSensitive="false" fromSubscription="false" />
    <match enabled="true" name="*10*.amazonaws.com*" pattern="*10*.amazonaws.com*" isRegex="false"
isBlackList="false" isMultiline="false" caseSensitive="false" fromSubscription="false" />
    <match enabled="true" name="*10*.compute*" pattern="*10*.compute*" isRegex="false" isBlackList="false"
isMultiline="false" caseSensitive="false" fromSubscription="false" />
    <match enabled="true" name="*.compute.internal*" pattern="*.compute.internal*" isRegex="false"
isBlackList="false" isMultiline="false" caseSensitive="false" fromSubscription="false" />
    <match enabled="true" name="*.ec2.internal*" pattern="*.ec2.internal*" isRegex="false" isBlackList="false"
isMultiline="false" caseSensitive="false" fromSubscription="false" />
    </matches>
    <manualconf host="localhost" port="8157" socksVersion="5" isSocks="true" username="" password="" domain="" />
    </proxy>
  </proxies>
</foxyproxy>
```

Notas:

- o a porta 8157 é a porta local utilizada para estabelecer o túnel SSH com o nó principal. Esse valor deve corresponder ao número da porta que você usou no PuTTY ou no terminal.
- o Os padrões de *ec2*.amazonaws.com* correspondem ao nome DNS público dos clusters na região us-east-1.
- o O padrão *ec2*.compute* corresponde ao nome DNS público dos clusters em todas as outras regiões.
- o O padrão 10* fornece acesso aos arquivos de log do JobTracker no Hadoop 1.x. Altere esse filtro se ele estiver em conflito com o plano de acesso da rede.

4. Clique no ícone FoxyProxy na barra de ferramentas e selecione Opções.

5. Clique em Importar/Exportar.

6. Clique em Escolher arquivo, selecione foxyproxy-settings.xml e clique em Abrir.

7. Na caixa de diálogo Importar configurações do FoxyProxy, clique em Adicionar.

8. Na página superior da página, para o Modo de proxy, escolha Usar Use proxies com base em padrões e prioridades predefinidos

9. Para abrir as interfaces da web, na barra de endereços do seu navegador, digite *master-public-dns* seguido do URL ou do número da porta.

Para obter uma lista completa de interfaces da web no nó principal, consulte [Visualizar interfaces da web hospedadas nos clusters do Amazon EMR](#).

Home x EMR – AWS Console x S3 Management Console x brunojardim/Jupyter-T x generate new key aws x Apache Livy x Apache Tez - Hortonw x +

https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1Y3QBHYPWEM17

Apps Deep Purple - Highw Gmail: e-mail do Goo Google Bandhub - Make Mus GCP Timesheet Projeto B3 Clusters Type AWS aws

aws Services Resource Groups IAM EC2 EMR Lambda S3 RDS CloudWatch Dynan

Amazon EMR

Clusters

Configuração de segurança

Sub-redes da VPC

Eventos

Ajuda

What's new

Clonar Encerrar Exportação de CLI da AWS

Cluster: EMR_BJ **Aguardando** Cluster ready after last step completed.

Resumo Histórico do aplicativo Monitoramento Hardware Eventos Etapas Configurações Ações de bootstrap

Conexões: Hue, Zeppelin, Servidor de histórico do Spark, Ganglia, HBase, JupyterHub, Gerenciador de recursos ... (Visualizar tudo)

DNS público principal ec2-35-169-124-144.compute-1.amazonaws.com SSH

Tags: Nome = BrunoJardim, Treinamento = BigData Visualizar todas/Editar

Resumo Detalhes da configuração Rede e hardware

Use proxies based on their pre-defined patterns and priorities

Use proxy emr-socks-proxy for all URLs

Use proxy Default for all URLs

Disable FoxyProxy

Options