

Objetivo

Este documento tem como objetivo de descrever o provisionamento do ambiente para prática de processamento de dados.

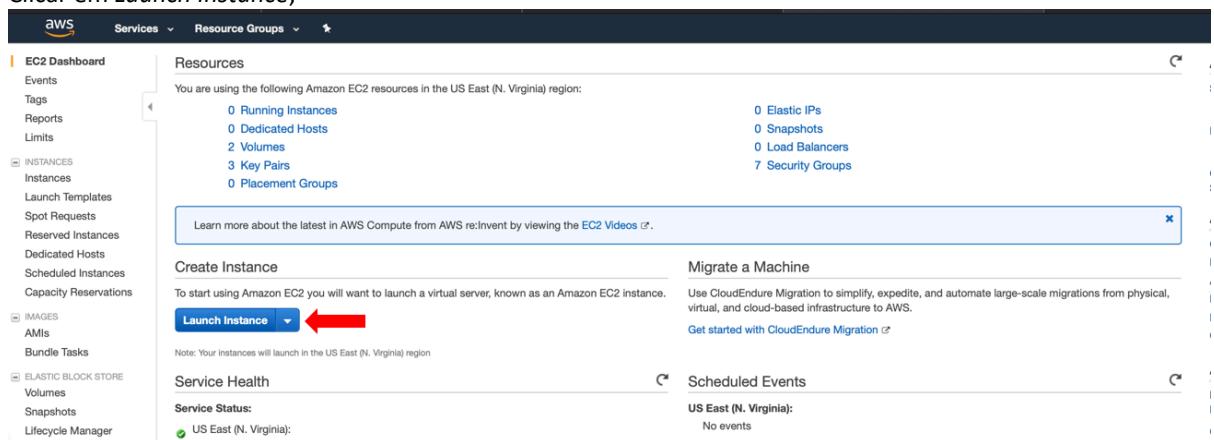
Atividades:

- Provisionamento Red Hat Enterprise Linux 8
 - StreamSets
- Provisionamento Windows Server 2019
 - SQL Server 2017
- Provisionamento Amazon Linux
 - Kafka
- Provisionamento EMR – Elastic MapReduce
 - Hive 2.3.5
 - Spark 2.4.4
 - Hue 4.4.0
 - Livy 0.6.0
 - Tez 0.9.2
 - Zeppelin 0.8.1
 - JupyterHub 1.0.0
- Configuração do ambiente para ingestão de dados ao Data Lake

Red Hat Enterprise Linux 8 – StreamSets

Provisionamento de instância EC2 (Elastic Compute Cloud):

1. Realizar *login* na console AWS;
2. Acessar *Services > Compute > EC2*;
3. Clicar em *Launch Instance*;



4. Selecionar ***Red Hat Enterprise Linux 8 (HVM), SSD Volume Type***

Step 1: Choose an Amazon Machine Image (AMI)
 An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.

Quick Start

- My AMIs
- AWS Marketplace
- Community AMIs
- Free tier only ①

AMI Name	Description	Root device type	Virtualization type	ENAs Enabled
Amazon Linux 2 AMI (HVM), SSD Volume Type - ami-0b69ea6bf7391e80 (64-bit x86) / ami-09c61c4850b7465cb (64-bit Arm)	Amazon Linux 2 comes with five years support. It provides Linux kernel 4.14 tuned for optimal performance on Amazon EC2, systemd 219, GCC 7.3, Glibc 2.26, Binutils 2.29.1, and the latest software packages through extras.	ebs	hvm	Yes
Amazon Linux AMI 2018.03.0 (HVM), SSD Volume Type - ami-00eb20669e0990cb4	The Amazon Linux AMI is an EBS-backed, AWS-supported image. The default image includes AWS command line tools, Python, Ruby, Perl, and Java. The repositories include Docker, PHP, MySQL, PostgreSQL, and other packages.	ebs	hvm	Yes
Red Hat Enterprise Linux 8 (HVM), SSD Volume Type - ami-0c322300a1dd5dc79 (64-bit x86) / ami-03587fa4048e0eb92 (64-bit Arm)	Red Hat Enterprise Linux version 8 (HVM), EBS General Purpose (SSD) Volume Type	ebs	hvm	Yes

Cancel and Exit

Select ② 64-bit (x86)
 Select ③ 64-bit (Arm)

5. Selecionar instância ***t2.xlarge*** e clicar em “***Next: Configure Instance Details***”

Step 2: Choose an Instance Type
 Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

Filter by: All instance types Current generation Show/Hide Columns

Currently selected: t2.xlarge (Variable ECUs, 4 vCPUs, 2.3 GHz, Intel Broadwell E5-2686v4, 16 GiB memory, EBS only)

Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance	IPv6 Support
General purpose	t2.nano	1	0.5	EBS only	-	Low to Moderate	Yes
General purpose	t2.micro Free tier eligible	1	1	EBS only	-	Low to Moderate	Yes
General purpose	t2.small	1	2	EBS only	-	Low to Moderate	Yes
General purpose	t2.medium	2	4	EBS only	-	Low to Moderate	Yes
General purpose	t2.large	2	8	EBS only	-	Low to Moderate	Yes
General purpose	t2.xlarge	4	16	EBS only	-	Moderate	Yes

6. Utilizar as configurações padrões e clicar em “***Next: Add Storage***”

Step 3: Configure Instance Details
 Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

Number of instances ④ 1 Launch into Auto Scaling Group ⑤

Purchasing option Request Spot Instances

Network vpc-a8f95d1 (default) Create new VPC
 Subnet No preference (default subnet in any Availability Zone) Create new subnet
 Auto-assign Public IP Use subnet setting (Enable)

Placement group Add instance to placement group
 Capacity Reservation Open Create new Capacity Reservation

IAM role None Create new IAM role

Shutdown behavior Stop
 Enable termination protection Protect against accidental termination
 Monitoring Enable CloudWatch detailed monitoring Additional charges apply.
 Tenancy Shared - Run a shared hardware instance Additional charges will apply for dedicated tenancy.
 Elastic Inference Add an Elastic Inference accelerator Additional charges apply.
 T2/T3 Unlimited Enable Additional charges may apply

Advanced Details

User data As text As file Input is already base64 encoded
 (Optional)

Cancel Previous Review and Launch Next: Add Storage

7. Alterar “**Size GiB**” para **30** e clicar em “**Next: Add Tags**”

Volume Type	Device	Snapshot	Size (GiB)	Volume Type	IOPS	Throughput (MB/s)	Delete on Termination	Encryption
Root	/dev/sda1	snap-0f9729669b35cf10f	30	General Purpose SSD (gp2)	100 / 3000	N/A	<input checked="" type="checkbox"/>	Not Encrypted

Add New Volume

Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. [Learn more](#) about free usage tier eligibility and usage restrictions.

Cancel Previous Review and Launch Next: Add Tags

8. Não adicionar tag e clicar em “**Next: Configure Security Group**”

9. Preencher nome, descrição, liberação para acesso externo da sua máquina local e clicar “**Review and Launch**”

Type	Protocol	Port Range	Source	Description
SSH	TCP	22	My IP 201.11.204.166/32	e.g. SSH for Admin Desktop
All traffic	All	0 - 65535	My IP 201.11.204.166/32	e.g. SSH for Admin Desktop

Add Rule Cancel Previous Review and Launch

Type	Protocol	Port Range	Source	Description
All traffic	All	0 - 65535	My IP 177.220.235.245/32	e.g. SSH for Admin Desktop

Add Rule Cancel Previous Review and Launch

10. Revisar as configurações e clicar em “**Launch**”

Step 7: Review Instance Launch

Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

AMI Details

Red Hat Enterprise Linux 8 (HVM), SSD Volume Type - ami-0c322300a1dd5dc79

Free tier eligible

Root Device Type: ebs Virtualization type: hvm

Instance Type

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t2.xlarge	Variable	4	16	EBS only	-	Moderate

Security Groups

Security group name: uniritter-aula
Description: Security Group para provisionamento de ambiente de teste

Type	Protocol	Port Range	Source	Description
SSH	TCP	22	201.11.204.166/32	

Cancel Previous Launch

11. Criar uma chave ou utilizar uma existente

- Informar nome
- Realizar download da chave
- Clicar em “**Launch Instances**”

Select an existing key pair or create a new key pair X

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store.

Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about [removing existing key pairs from a public AMI](#).

Create a new key pair

Key pair name

uniritter

Download Key Pair

You have to download the **private key file** (*.pem file) before you can continue.
Store it in a secure and accessible location. You will not be able to download the file again after it's created.

Cancel Launch Instances

12. Clicar no ID da instância e acompanhar a inicialização da máquina

The screenshot shows the AWS EC2 Dashboard with the Instances page selected. The search bar at the top has the value 'search : i-09af568c45e3ab33b'. The main table lists one instance named 'UNIRITTER-STREAMSETS' with the following details:

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status
UNIRITTER-STREAMSETS	i-09af568c45e3ab33b	t2.xlarge	us-east-1c	running	2/2 checks...	None

13. Ajustar as permissões da chave para acesso via SSH

```
sh-3.2# pwd
/Users/roberto.silva/Documents/keys/uniritter
sh-3.2#
sh-3.2# ls -l
total 8
-rw-r--r--@ 1 roberto.silva 894878024 1692 Oct 16 08:38 uniritter.pem
sh-3.2#
sh-3.2#
sh-3.2# chmod 400 uniritter.pem
sh-3.2#
sh-3.2#
sh-3.2# ls -l
total 8
-r-----@ 1 roberto.silva 894878024 1692 Oct 16 08:38 uniritter.pem
sh-3.2#
```

14. Acessar a instância provisionada

```
sh-3.2#
sh-3.2# ssh -i uniritter.pem ec2-user@34.201.140.253
The authenticity of host '34.201.140.253 (34.201.140.253)' can't be established.
ECDSA key fingerprint is SHA256:iXcwSoKx9LkHchZ6AL4E6VE+vulcOMiu8QByYm+81K8.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '34.201.140.253' (ECDSA) to the list of known hosts.
[ec2-user@ip-172-31-46-168 ~]$
```

15. Instalar o Java

```
[ec2-user@ip-172-31-46-168 ~]$
[ec2-user@ip-172-31-46-168 ~]$ sudo su
[root@ip-172-31-46-168 ec2-user]#
[root@ip-172-31-46-168 ec2-user]# yum -y install java-1.8.0-openjdk
Last metadata expiration check: 0:10:20 ago on Wed 16 Oct 2019 11:51:51 AM UTC.
Dependencies resolved.
=====
Package                               Arch      Version
=====
Installing:
  java-1.8.0-openjdk                  x86_64    1:1.8.0.222.b1
Installing dependencies:
  libdatrie                            x86_64    0.2.9-7.el8
  giflib                               x86_64    5.1.4-3.el8
  libXext                               x86_64    1.3.3-9.el8
  libXrandr                            x86_64    1.5.1-7.el8
  graphite2                            x86_64    1.3.10-10.el8
```

16. Verificar a instalação do Java

```
[root@ip-172-31-46-168 ec2-user]# java -version
[root@ip-172-31-46-168 ec2-user]# java -version
openjdk version "1.8.0_222"
OpenJDK Runtime Environment (build 1.8.0_222-b10)
OpenJDK 64-Bit Server VM (build 25.222-b10, mixed mode)
[root@ip-172-31-46-168 ec2-user]#
```

17. Instalar o WGET

```
[root@ip-172-31-46-168 ec2-user]# yum install wget
[root@ip-172-31-46-168 ec2-user]# yum install wget
Last metadata expiration check: 0:33:29 ago on Wed 16 Oct 2019 11:51:51 AM UTC.
Dependencies resolved.

=====
Package                               Arch          Version
=====
Installing:
  wget                                x86_64      1.19.5-7.el8_0.1

Transaction Summary
=====
Install 1 Package

Total download size: 734 k
Installed size: 2.8 M
Is this ok [Y/N]: y
Downloading Packages:
wget-1.19.5-7.el8_0.1.x86_64.rpm
-----
Total
Running transaction check
Transaction check succeeded.
Running transaction test
Transaction test succeeded.
Running transaction
  Preparing       :
  Installing     : wget-1.19.5-7.el8_0.1.x86_64
  Running scriptlet: wget-1.19.5-7.el8_0.1.x86_64
  Verifying      : wget-1.19.5-7.el8_0.1.x86_64

Installed:
  wget-1.19.5-7.el8_0.1.x86_64

Complete!
```

18. Baixar o StreamSets

Comando: wget <https://s3-us-west-2.amazonaws.com/archives.streamsets.com/datacollector/3.11.0/rpm/el7/streamsets-datacollector-3.11.0-el7-all-rpms.tar>

<https://s3-us-west-2.amazonaws.com/archives.streamsets.com/datacollector/3.11.0/rpm/el7/streamsets-datacollector-3.11.0-el7-all-rpms.tar>

```
[root@ip-172-31-46-168 ec2-user]#
[root@ip-172-31-46-168 ec2-user]# wget https://s3-us-west-2.amazonaws.com/archives.streamsets.com/datacollector/3.11.0/rpm/el7/streamsets-datacollector-3.11.0-el7-all-rpms.tar
--2019-10-16 12:26:26--  https://s3-us-west-2.amazonaws.com/archives.streamsets.com/datacollector/3.11.0/rpm/el7/streamsets-datacollector-3.11.0-el7-all-rpms.tar
Resolving s3-us-west-2.amazonaws.com (s3-us-west-2.amazonaws.com) ... 52.218.220.56
Connecting to s3-us-west-2.amazonaws.com (s3-us-west-2.amazonaws.com)|52.218.220.56|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 5544171520 (5.2G) [application/x-tar]
Saving to: 'streamsets-datacollector-3.11.0-el7-all-rpms.tar'

streamsets-datacollector-3.11.0-el7-all-rpms.tar    7%[=====]>
```

19. Descompactar o StreamSets

Comando: tar -xvf streamsets-datacollector-3.11.0-el7-all-rpms.tar

```
[root@ip-172-31-46-168 ec2-user]#  
[root@ip-172-31-46-168 ec2-user]# tar -xvf streamsets-datacollector-3.11.0-el7-all-rpms.tar  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-aerospike-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-kafka_1_0-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-kafka_1_1-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-kafka_2_0-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-kudu_1_3-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-kudu_1_4-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-kudu_1_5-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-kudu_1_6-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-kudu_1_7-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-pulsar_2-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-apache-solr_6_1_0-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-aws-lib-3.11.0-1.noarch.rpm  
streamsets-datacollector-3.11.0-el7-all-rpms/streamsets-datacollector-aws-secrets-manager-credentialstore-lib-3
```

20. Instalar o StreamSets

Comando: yum localinstall streamsets*.rpm

```
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]#  
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]#  
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]# pwd  
/home/ec2-user/streamsets-datacollector-3.11.0-el7-all-rpms  
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]# yum localinstall streamsets*.rpm  
Last metadata expiration check: 0:44:38 ago on Wed 16 Oct 2019 11:51:51 AM UTC.  
Dependencies resolved.  
=====  
           Package                                Arch  
=====  
Installing:  
  streamsets-datacollector                         noarch  
  streamsets-datacollector-aerospike-lib          noarch  
  streamsets-datacollector-apache-kafka_1_0-lib    noarch  
  streamsets-datacollector-apache-kafka_1_1-lib    noarch  
  streamsets-datacollector-apache-kafka_2_0-lib    noarch  
  streamsets-datacollector-apache-kudu_1_3-lib     noarch
```

21. Alterar o owner do diretório de instalação

Comando:

chown -R sdc:sdc /etc/sdc/

chown -R sdc:sdc /opt/streamsets-datacollector/

```
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]#  
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]#  
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]#  
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]# chown -R sdc:sdc /etc/sdc/  
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]# chown -R sdc:sdc /opt/streamsets-datacollector/  
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]#
```

22. Iniciar o StreamSets

```
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]# systemctl start sdc
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]# systemctl status sdc
● sdc.service - StreamSets Data Collector (SDC)
   Loaded: loaded (/usr/lib/systemd/system/sdc.service; disabled; vendor preset: disabled)
   Active: active (running) since Wed 2019-10-16 12:43:01 UTC; 12s ago
     Main PID: 14504 (_sdc)
        Tasks: 19 (limit: 26213)
       Memory: 680.5M
      CGroup: /system.slice/sdc.service
              └─14504 /bin/bash /opt/streamsets-datacollector/libexec/_sdc -verbose
                  ├─14550 /usr/bin/java -classpath /opt/streamsets-datacollector/libexec/bootstrap-libs/main/streamsets-datacollector-bootstrap-3.11.0.jar:
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: BOOTSTRAP_CLASSPATH          : /opt/streamsets-datacollector/libexec/bootstrap-
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: API_CLASSPATH                 : /opt/streamsets-datacollector/api-lib/*.jar
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: CONTAINER_CLASSPATH           : /etc/sdc:/opt/streamsets-datacollector/container
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: LIBS_COMMON_LIB_DIR         : /opt/streamsets-datacollector/libs-common-lib
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: STREAMSETS_LIBRARIES_DIR    : /opt/streamsets-datacollector/streamsets-libs
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: STREAMSETS_LIBRARIES_EXTRA_DIR : /opt/streamsets-datacollector/streamsets-libs-ex
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: USER_LIBRARIES_DIR          : /opt/streamsets-datacollector/user-libs
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: JAVA_OPTS                   : -Djava.security.manager -Djava.security.policy=f
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: MAIN CLASS                : com.streamsets.datacollector.main.DataCollectorM
Oct 16 12:43:03 ip-172-31-46-168.ec2.internal streamsets[14504]: Logging initialized @1390ms to org.eclipse.jetty.util.log.Slf4jLog
lines 1-20/20 (END)
```

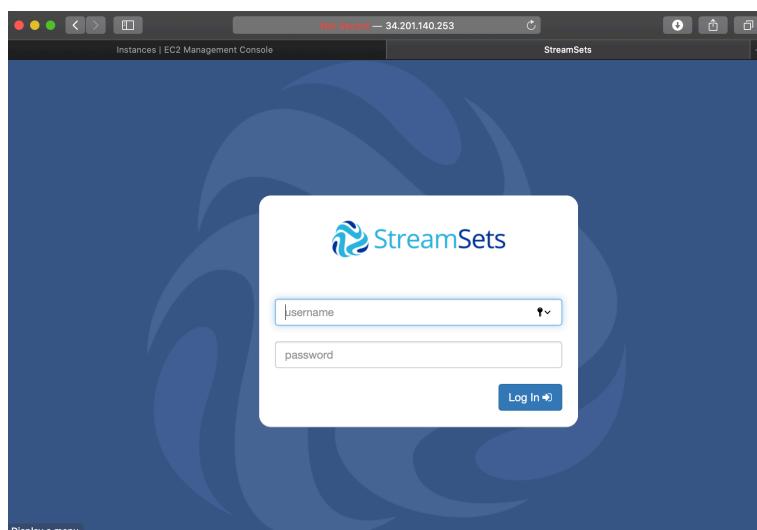
Auardar o serviço subir.

```
[root@ip-172-31-46-168 streamsets-datacollector-3.11.0-el7-all-rpms]# systemctl status sdc
● sdc.service - StreamSets Data Collector (SDC)
   Loaded: loaded (/usr/lib/systemd/system/sdc.service; disabled; vendor preset: disabled)
   Active: active (running) since Wed 2019-10-16 12:43:01 UTC; 1min 4s ago
     Main PID: 14504 (_sdc)
        Tasks: 33 (limit: 26213)
       Memory: 926.0M
      CGroup: /system.slice/sdc.service
              └─14504 /bin/bash /opt/streamsets-datacollector/libexec/_sdc -verbose
                  ├─14550 /usr/bin/java -classpath /opt/streamsets-datacollector/libexec/bootstrap-libs/main/streamsets-datacollector-bootstrap-3.

Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: API_CLASSPATH          : /opt/streamsets-datacollector/api-lib/*
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: CONTAINER_CLASSPATH           : /etc/sdc:/opt/streamsets-datacollector/
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: LIBS_COMMON_LIB_DIR         : /opt/streamsets-datacollector/libs-comm
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: STREAMSETS_LIBRARIES_DIR    : /opt/streamsets-datacollector/streamset
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: STREAMSETS_LIBRARIES_EXTRA_DIR : /opt/streamsets-datacollector/streamset
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: USER_LIBRARIES_DIR          : /opt/streamsets-datacollector/user-libs
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: JAVA_OPTS                   : -Djava.security.manager -Djava.security.pol
Oct 16 12:43:01 ip-172-31-46-168.ec2.internal streamsets[14504]: MAIN CLASS                : com.streamsets.datacollector.main.DataCol
Oct 16 12:43:03 ip-172-31-46-168.ec2.internal streamsets[14504]: Logging initialized @1390ms to org.eclipse.jetty.util.log.Slf4jLog
Oct 16 12:43:26 ip-172-31-46-168.ec2.internal streamsets[14504]: Running on URI : 'http://ip-172-31-46-168.ec2.internal:18630'
lines 1-20/20 (END)
```

- Acessar a console do StreamSets

Utilizar o IP publico.



Windows Server 2019 – SQL Server 2017

Provisionamento de instância EC2 (Elastic Computer Cloud):

1. Realizar *login* na console AWS;
2. Acessar *Services > Compute > EC2*;
3. Clicar em *Launch Instance*;

The screenshot shows the AWS EC2 Dashboard. On the left, there's a sidebar with categories like EC2 Dashboard, Instances, Images, and Elastic Block Store. The main area shows 'Resources' with counts for Running Instances, Dedicated Hosts, Volumes, Key Pairs, and Placement Groups. Below this is a 'Create Instance' section with a 'Launch Instance' button highlighted by a red arrow. To the right is a 'Migrate a Machine' section with migration statistics.

4. Selecionar Microsoft Windows Server 2019

The screenshot shows the 'Choose an Amazon Machine Image (AMI)' step of the AWS instance creation wizard. It lists three Windows Server 2019 AMIs: 'Microsoft Windows Server 2019 Base - ami-0d4df21ffeb914d61', 'Microsoft Windows Server 2019 Base with Containers - ami-03fb9bcb97e9cabb', and 'Microsoft Windows Server 2019 with SQL Server 2017 Standard - ami-00da65e71856d58bb'. Each item has a 'Select' button to its right, with a red arrow pointing to the first one. The sidebar on the left shows 'Quick Start (6)' with options for My AMIs, AWS Marketplace, and Community AMIs, and a 'Free tier only' checkbox.

5. Selecionar o tipo de máquina t2.xlarge e clicar em “**Next: Configure Instance Details**”

[1. Choose AMI](#) [2. Choose Instance Type](#) [3. Configure Instance](#) [4. Add Storage](#) [5. Add Tags](#) [6. Configure Security Group](#) [7. Review](#)

Step 2: Choose an Instance Type

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

Filter by: [All instance types](#) [Current generation](#) [Show/Hide Columns](#)

Currently selected: t2.xlarge (Variable ECUs, 4 vCPUs, 2.3 GHz, Intel Broadwell E5-2686v4, 16 GiB memory, EBS only)

	Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance	IPv6 Support
<input type="checkbox"/>	General purpose	t2.nano	1	0.5	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	General purpose	t2.micro <small>Free tier eligible</small>	1	1	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	General purpose	t2.small	1	2	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	General purpose	t2.medium	2	4	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	General purpose	t2.large	2	8	EBS only	-	Low to Moderate	Yes
<input checked="" type="checkbox"/>	General purpose	t2.xlarge	4	16	EBS only	-	Moderate	Yes
<input type="checkbox"/>	General purpose	t2.2xlarge	8	32	EBS only	-	Moderate	Yes

6. Utilizar as configurações padrões e clicar em “**Next: Add Storage**”

[1. Choose AMI](#) [2. Choose Instance Type](#) [3. Configure Instance](#) [4. Add Storage](#) [5. Add Tags](#) [6. Configure Security Group](#) [7. Review](#)

Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

Number of instances <small>(i)</small>	<input type="text" value="1"/> Launch into Auto Scaling Group <small>(i)</small>
Purchasing option <small>(i)</small>	<input type="checkbox"/> Request Spot instances
Network <small>(i)</small>	<input type="text" value="vpc-aa8f95d1 (default)"/> <small>C</small> Create new VPC
Subnet <small>(i)</small>	<input type="text" value="No preference (default subnet in any Availability zone)"/> <small>C</small> Create new subnet
Auto-assign Public IP <small>(i)</small>	<input type="text" value="Use subnet setting (Enable)"/>
Placement group <small>(i)</small>	<input type="checkbox"/> Add instance to placement group
Capacity Reservation <small>(i)</small>	<input type="text" value="Open"/> <small>C</small> Create new Capacity Reservation
Domain join directory <small>(i)</small>	<input type="text" value="No directory"/> <small>C</small> Create new directory
IAM role <small>(i)</small>	<input type="text" value="None"/> <small>C</small> Create new IAM role
Shutdown behavior <small>(i)</small>	<input type="text" value="Stop"/>
Enable termination protection <small>(i)</small>	<input type="checkbox"/> Protect against accidental termination
Monitoring <small>(i)</small>	<input type="checkbox"/> Enable CloudWatch detailed monitoring <small>Additional charges apply.</small>
Tenancy <small>(i)</small>	<input type="text" value="Shared - Run a shared hardware instance"/> <small>Additional charges will apply for dedicated tenancy.</small>
Elastic Graphics <small>(i)</small>	<input type="checkbox"/> Add Graphics Acceleration <small>Additional charges apply.</small>
T2/T3 Unlimited <small>(i)</small>	<input type="checkbox"/> Enable <small>Additional charges may apply</small>
Advanced Details	
User data <small>(i)</small>	<input checked="" type="radio"/> As text <input type="radio"/> As file <input type="checkbox"/> Input is already base64 encoded <small>(Optional)</small>

[Cancel](#) [Previous](#) [Review and Launch](#) [Next: Add Storage](#)

Disciplina: Processamento de Grandes Volumes de Dados

Professor: Roberto Galvão

7. Ajustar o tamanho do disco se necessário para volumetria do teste e clicar em “**Next: Add Tags**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 4: Add Storage

Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. [Learn more](#) about storage options in Amazon EC2.

Volume Type	Device	Snapshot	Size (GiB)	Volume Type	IOPS	Throughput (MB/s)	Delete on Termination	Encryption
Root	/dev/sda1	snap-000d7a18e471fcd89	50	General Purpose SSD (gp2)	150 / 3000	N/A	<input checked="" type="checkbox"/>	Not Encrypted
Add New Volume								

Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. [Learn more](#) about free usage tier eligibility and usage restrictions.

Cancel Previous **Review and Launch** Next: Add Tags

8. Não utilizaremos Tags. Clicar em “**Next Configure Security Group**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 5: Add Tags

A tag consists of a case-sensitive key-value pair. For example, you could define a tag with key = Name and value = Webserver. A copy of a tag can be applied to volumes, instances or both. Tags will be applied to all instances and volumes. [Learn more](#) about tagging your Amazon EC2 resources.

Key (128 characters maximum)	Value (256 characters maximum)	Instances (1) Volumes (1)
This resource currently has no tags		
Choose the Add tag button or click to add a Name tag . Make sure your IAM policy includes permissions to create tags.		
Add Tag	(Up to 50 tags maximum)	

Cancel Previous **Review and Launch** Next: Configure Security Group

Disciplina: Processamento de Grandes Volumes de Dados
Professor: Roberto Galvão

9. Selecionar o *Security Group* existente que criamos no provisionamento anterior e clicar em “**Review and Launch**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group **7. Review**

Step 6: Configure Security Group
A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. [Learn more](#) about Amazon EC2 security groups.

Assign a security group: Create a new security group Select an existing security group

Security Group ID	Name	Description	Actions
sg-0f5dc2bb4c60ec6df	uniritter-aula	Security Group para provisionamento de ambiente de teste	Copy to new

Inbound rules for sg-0f5dc2bb4c60ec6df (Selected security groups: sg-0f5dc2bb4c60ec6df)

Type	Protocol	Port Range	Source	Description
All traffic	All	All	201.37.162.34/32	
SSH	TCP	22	201.37.162.34/32	

[Cancel](#) [Previous](#) **Review and Launch**

10. Revisar as configurações e clicar em “**Launch**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group **7. Review**

Step 7: Review Instance Launch
Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

AMI Details [Edit AMI](#)

	Microsoft Windows Server 2019 Base - ami-0d4df21ffeb914d61
<small>Free tier eligible</small>	Microsoft Windows 2019 Datacenter edition. [English]
	Root Device Type: ebs Virtualization type: hvm

If you plan to use this AMI for an application that benefits from Microsoft License Mobility, fill out the [License Mobility Form](#). Don't show me this again

Instance Type [Edit instance type](#)

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t2.xlarge	Variable	4	16	EBS only	-	Moderate

Security Groups [Edit security groups](#)

Security Group ID	Name	Description
sg-0f5dc2bb4c60ec6df	uniritter-aula	Security Group para provisionamento de ambiente de teste

All selected security groups inbound rules

Type	Protocol	Port Range	Source	Description
All traffic	All	All	201.37.162.34/32	
SSH	TCP	22	201.37.162.34/32	

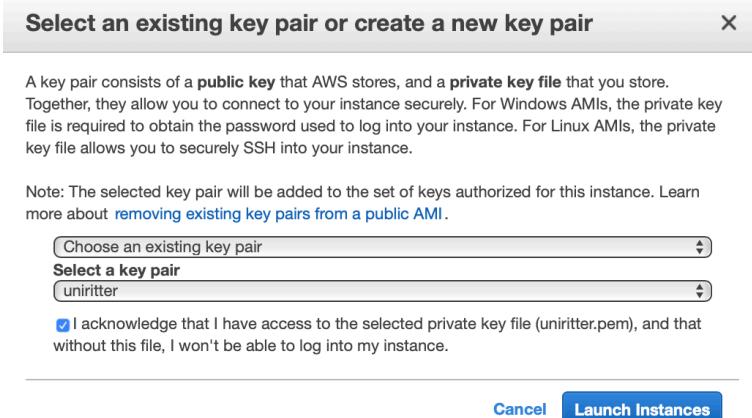
Instance Details [Edit instance details](#)

Storage [Edit storage](#)

Tags [Edit tags](#)

[Cancel](#) [Previous](#) **Launch**

11. Selecionar a chave que criamos no provisionamento anterior e clicar em “**Launch**”



12. Clicar no ID da instância e aguardar o provisionamento da máquina

Launch Status

The screenshot shows a "Your instances are now launching" message with a red arrow pointing to the instance ID "i-03a479b837779fcd8". Below it is a section for estimated charges with a blue info icon. The URL at the top is "https://aws.amazon.com/pt/billing/estimated-charges/".

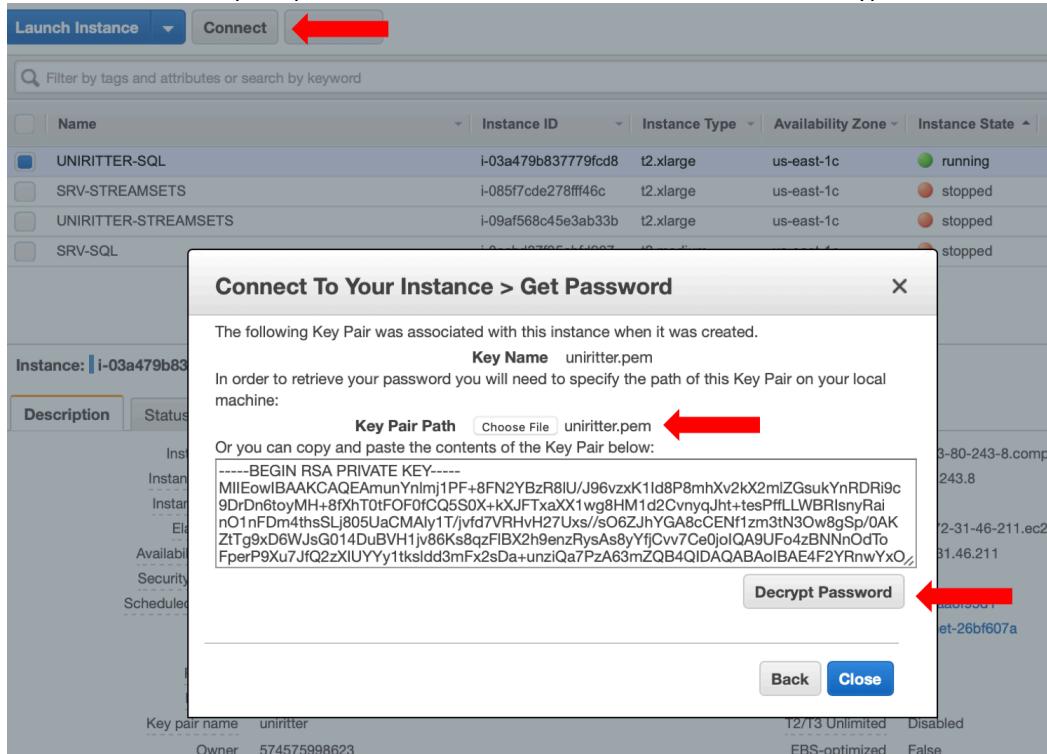
How to connect to your instances

Your instances are launching, and it may take a few minutes until they are in the running state, when they will be ready for you to use. Use the following steps to connect to your instances.

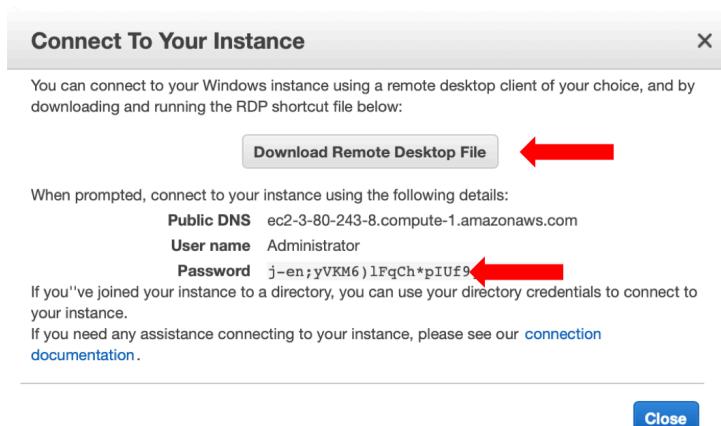
The screenshot shows the AWS EC2 Dashboard with the "Instances" tab selected. The left sidebar includes "EC2 Dashboard", "Events", "Tags", "Reports", "Limits", and "INSTANCES" with sub-options "Instances", "Launch Templates", "Spot Requests", and "Reserved Instances". The main area displays a table of instances:

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	
UNIRITTER-SQL	i-03a479b837779fcd8	t2.xlarge	us-east-1c	running	Initializing	/
SRV-STREAMSETS	i-085f7cde278fff46c	t2.xlarge	us-east-1c	stopped		/
UNIRITTER-STREAMSETS	i-09af568c45e3ab33b	t2.xlarge	us-east-1c	stopped		/
SRV-SQL	i-0ecbd37f05abfd907	t2.medium	us-east-1c	stopped		/

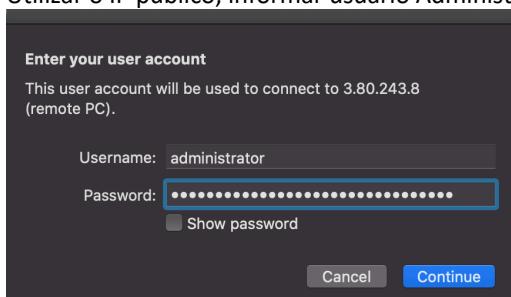
13. Selecionar a máquina provisionada, clicar em Connect, Choose File e Decrypt Password.



14. Copiar o Password e se necessário realize o download do Remote Desktop para acesso remoto ao servidor



15. Utilizar o IP publico, informar usuário Administrator e o Password copiado anteriormente



16. Após login ao servidor acessar Server Manager, Local Server e clicar em ON no marcado abaixo.

The screenshot shows the 'Server Manager' interface with 'Local Server' selected. In the 'PROPERTIES' section, under 'Windows Defender Firewall', there is a checkbox labeled 'Real-Time Protection: On'. This checkbox is checked (indicated by a green checkmark) and is highlighted with a red arrow.

17. Alterar para OFF em Administrators e reiniciar o servidor.

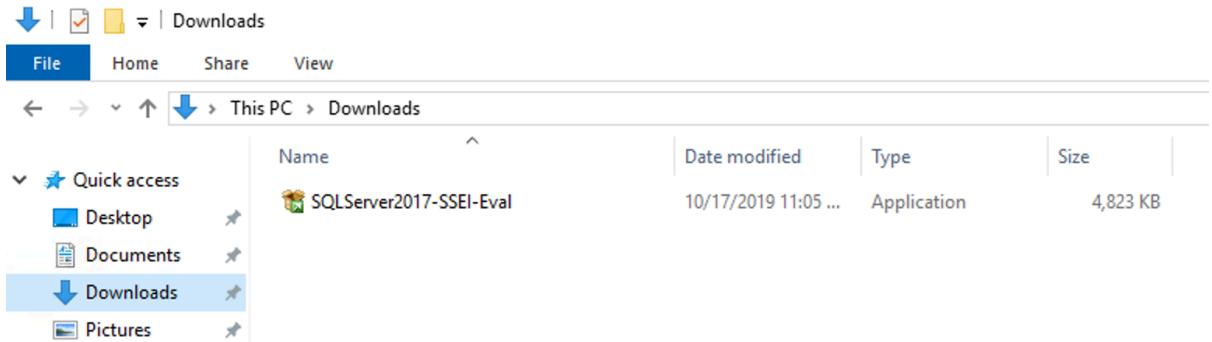
The screenshot shows the 'Internet Explorer Enhanced Security Configuration' dialog box. It has two main sections: 'Administrators:' and 'Users:'. Under both sections, there are two radio button options: 'On (Recommended)' (selected and highlighted with a red arrow) and 'Off'. Below the sections is a link 'More about Internet Explorer Enhanced Security Configuration'.

18. Realizar download do SQL Server Trial

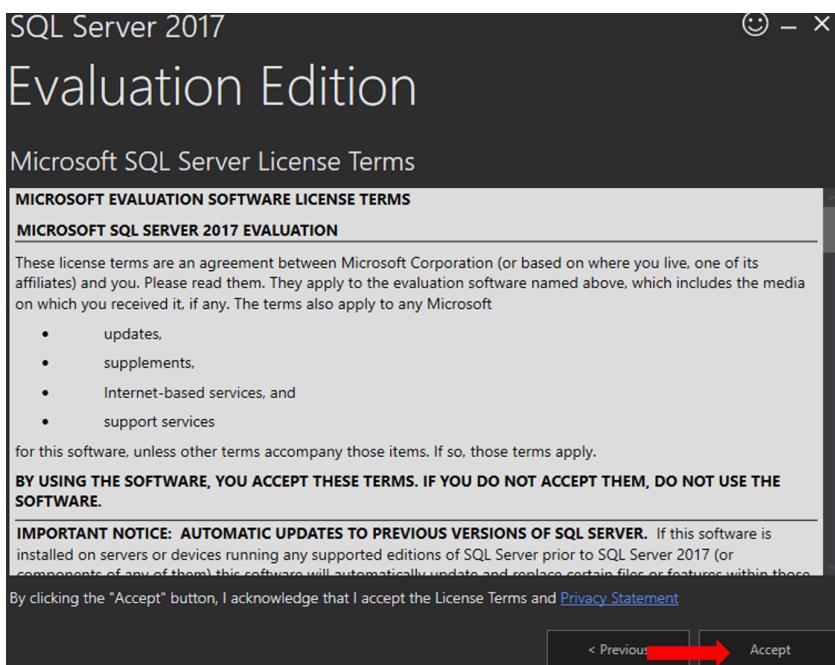
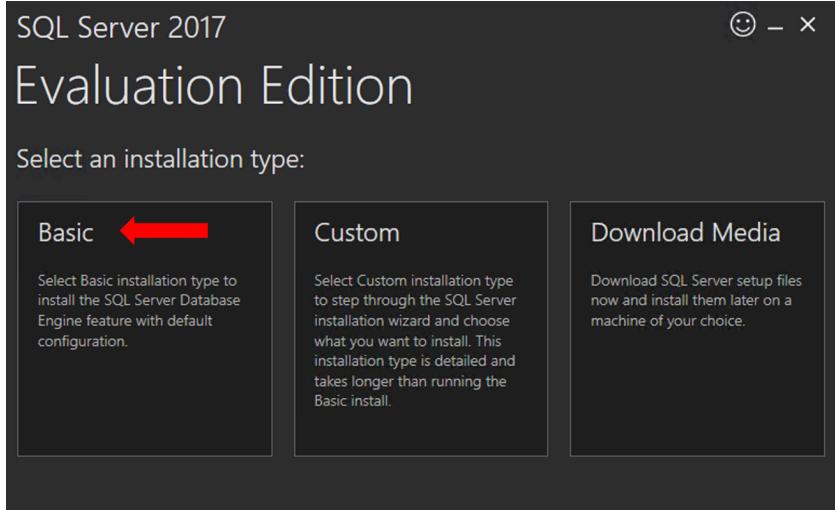
Link: <https://www.microsoft.com/en-us/sql-server/sql-server-downloads>

The screenshot shows the Microsoft Data platform website with a banner 'Try SQL Server on-premises or in the cloud'. It features two main sections: 'SQL Server 2017 on-premises' (dark blue background) and 'SQL Server in the cloud' (light blue background). The 'on-premises' section includes a 'Download free trial' button, which is highlighted with a red arrow. The 'in the cloud' section includes a 'Start free >' button.

19. Executar o instalador do SQL Server

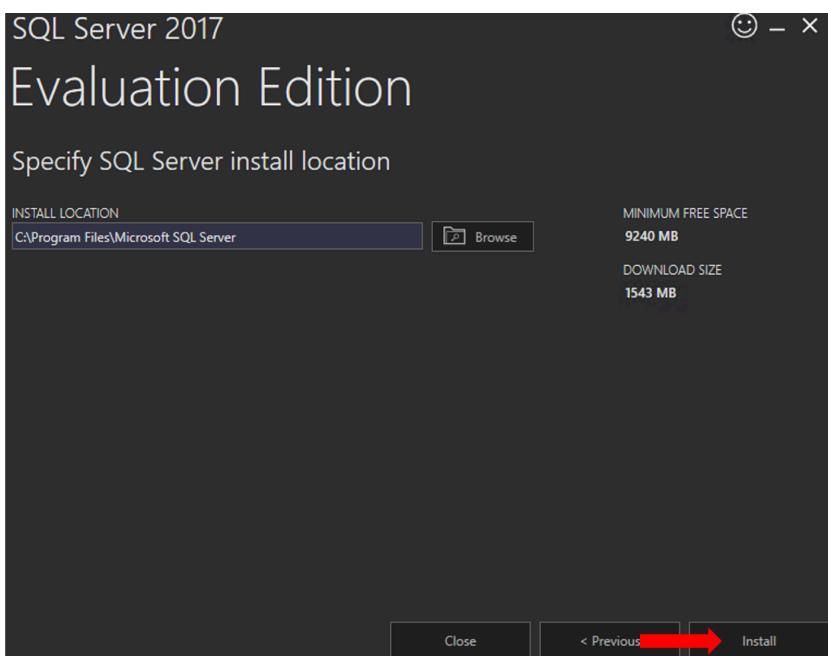


20. Realizar a instalação e configuração do SQL Server conforme imagens abaixo.

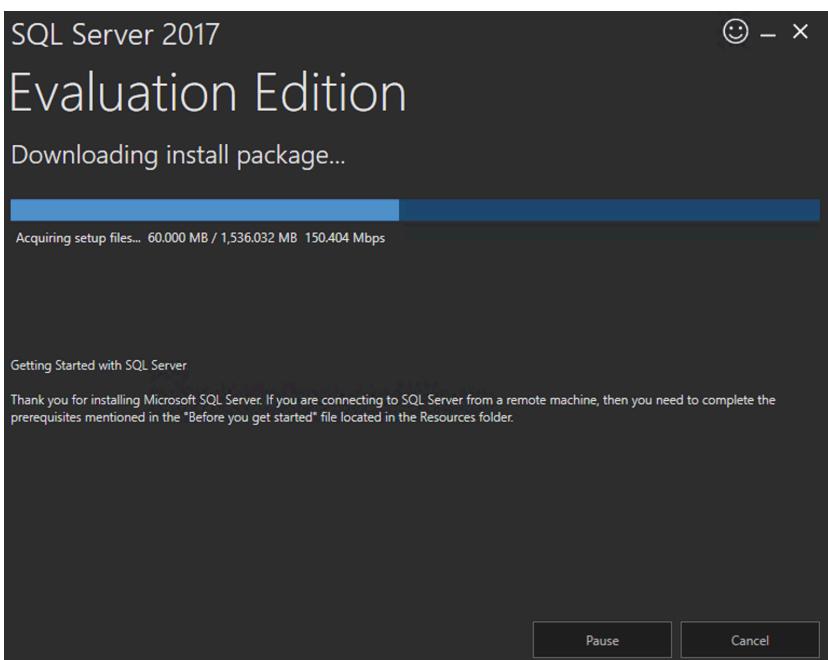


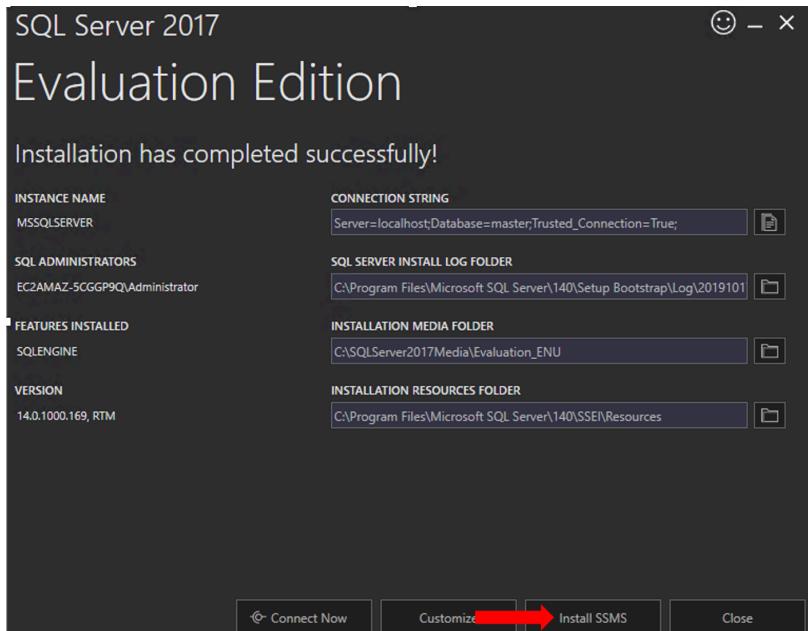
Disciplina: Processamento de Grandes Volumes de Dados

Professor: Roberto Galvão



*** Aguardar o download





[Microsoft](#) | [SQL Docs](#) Overview ▾ Install ▾ Secure ▾ Develop ▾ More ▾ [Download SQL Server](#)

Docs / SQL / Tools / SQL Server Management Studio (SSMS) / Download SSMS

Version
SQL Server 2019
Filter by title

Download SSMS

Release Notes

> Overview

> Tutorials

Download SQL Server Management Studio (SSMS)

10/03/2019 • 4 minutes to read • 22

In this article

Download SSMS 18.3.1 ← (highlighted with a red arrow)
Available languages (SSMS 18.3.1)
New in this release (SSMS 18.3.1)
Supported SQL offerings (SSMS 18.3.1)

RELEASE 18.3.1
Microsoft SQL Server Management Studio

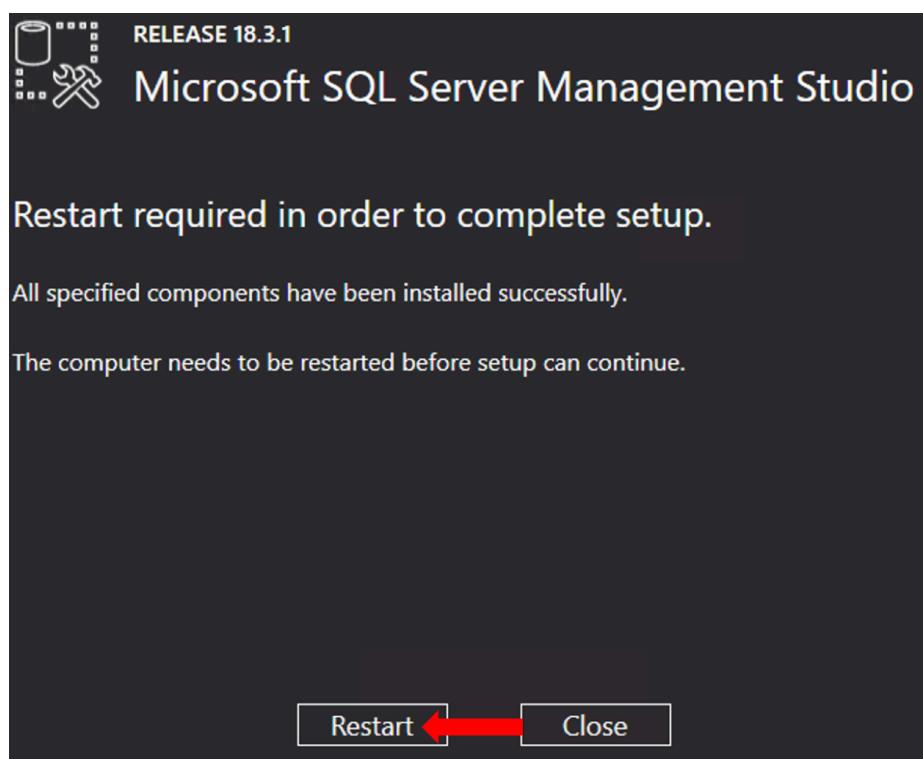
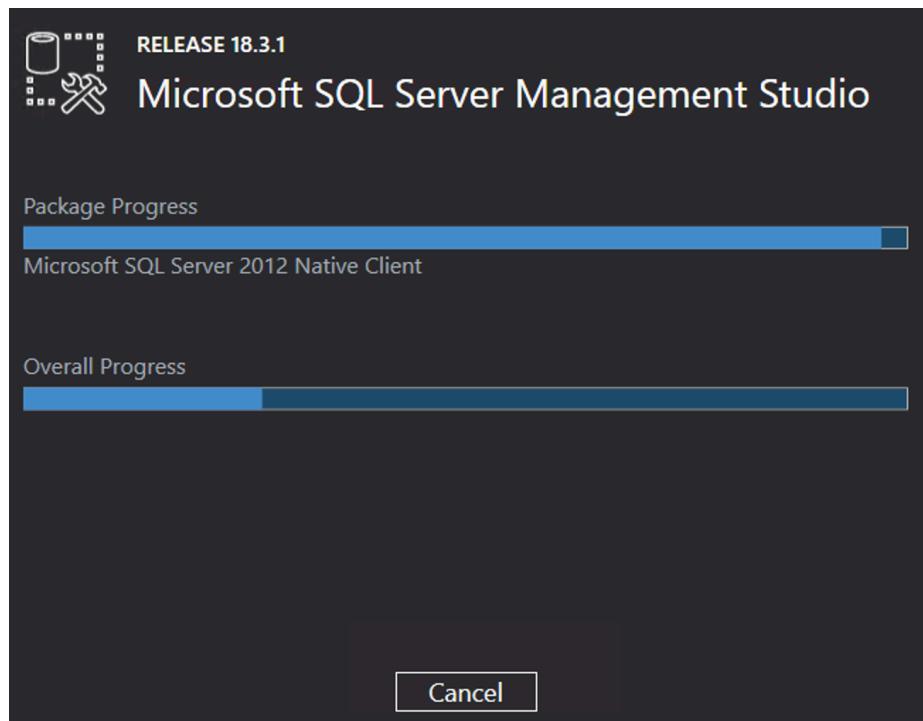
Welcome. Click "Install" to begin.

Location:
C:\Program Files (x86)\Microsoft SQL Server Management Studio 18 [Change](#)

By clicking the "Install" button, I acknowledge that I accept the [License Terms](#) and [Privacy Statement](#).

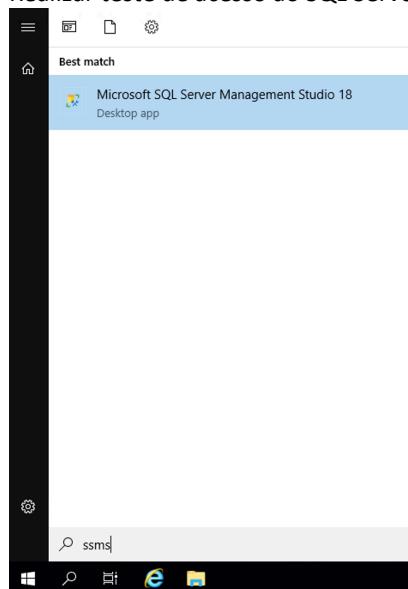
SQL Server Management Studio transmits information about your installation experience, as well as other usage and performance data, to Microsoft to help improve the product. To learn more about data processing and privacy controls, and to turn off the collection of this information after installation, see the [documentation](#).

Install ← (highlighted with a red arrow) Close

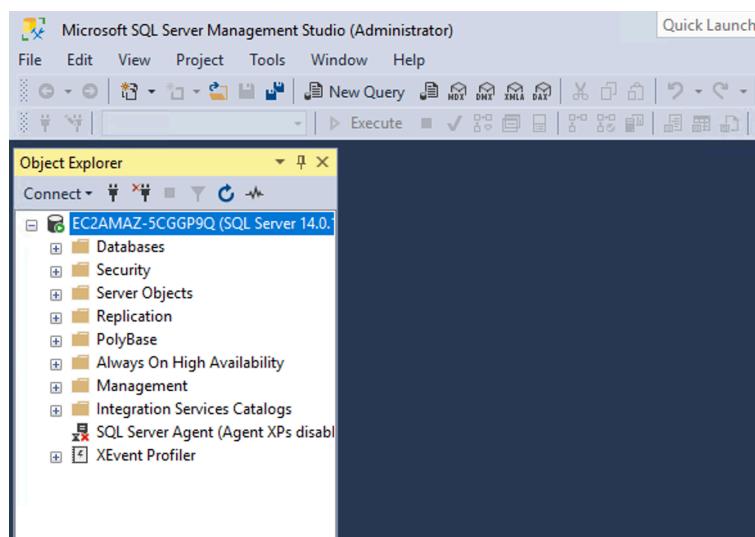
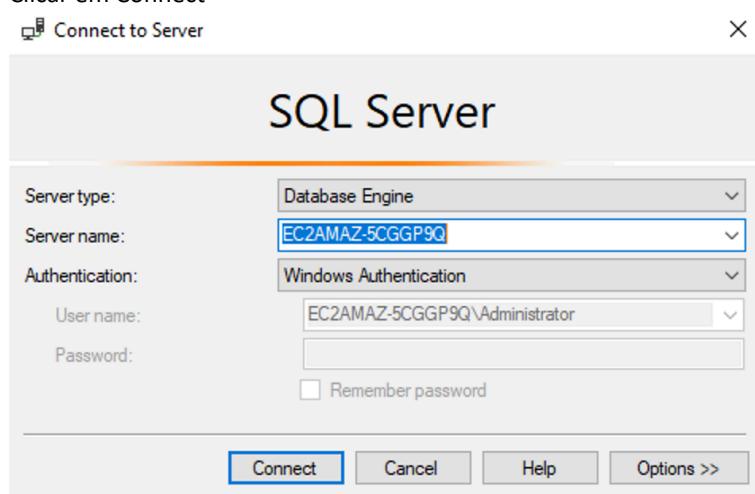


21. Desativar o Firewall do Windows Server.

22. Realizar teste de acesso ao SQL Server com SSMS



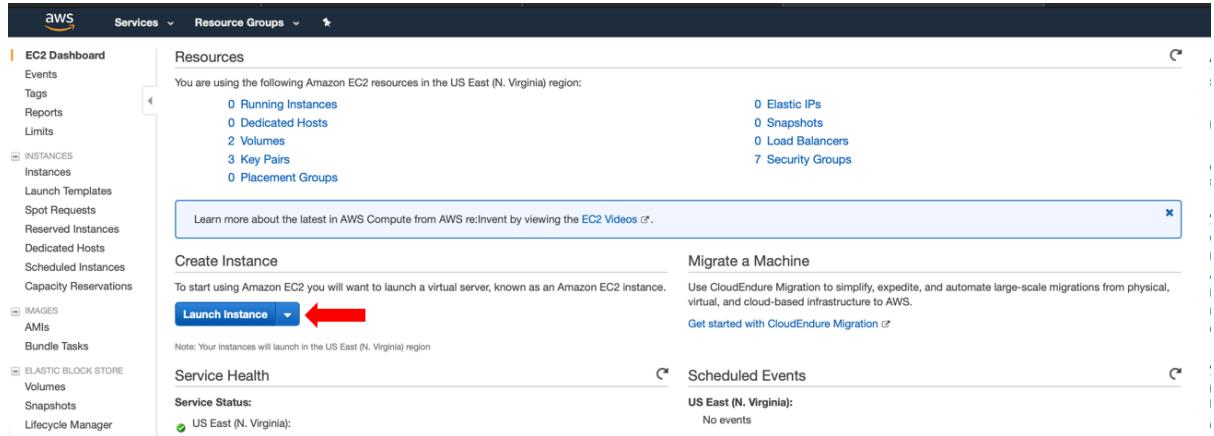
23. Clicar em Connect



Amazon Linux 2 – Kafka

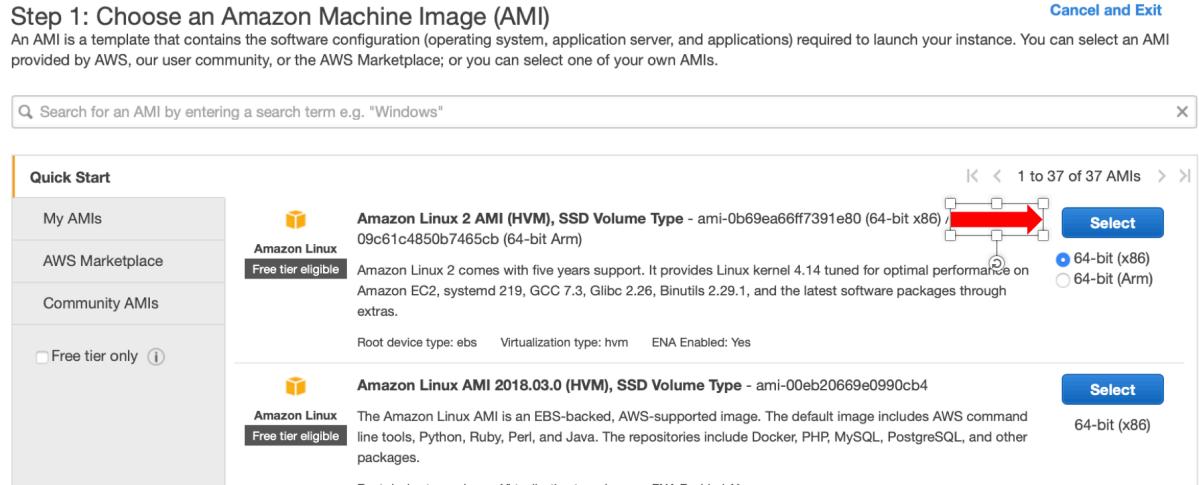
Provisionamento de instância EC2 (Elastic Computer Cloud):

1. Realizar *login* na console AWS;
2. Acessar *Services > Compute > EC2*;
3. Clicar em *Launch Instance*;



4. Seleciona Amazon Linux 2

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review



5. Selecionar o tipo de máquina t2.medium e clicar em “**Next: Configure Instance Details**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 2: Choose an Instance Type

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

Filter by: All instance types ▾ Current generation ▾ Show/Hide Columns

Currently selected: t2.medium (Variable ECUs, 2 vCPUs, 2.3 GHz, Intel Broadwell E5-2686v4, 4 GiB memory, EBS only)

	Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance	IPv6 Support
<input type="checkbox"/>	General purpose	t2.nano	1	0.5	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	General purpose	t2.micro Free tier eligible	1	1	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	General purpose	t2.small	1	2	EBS only	-	Low to Moderate	Yes
<input checked="" type="checkbox"/>	General purpose	t2.medium	2	4	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	General purpose	t2.large	2	8	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	General purpose	t2.xlarge	4	16	EBS only	-	Moderate	Yes
<input type="checkbox"/>	General purpose	t2.2xlarge	8	32	EBS only	-	Moderate	Yes

Cancel Previous Review and Launch Next: Configure Instance Details

6. Utilizar as configurações padrões e clicar em “**Next: Add Storage**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

Number of instances	<input type="text" value="1"/>	Launch into Auto Scaling Group (i)
Purchasing option	<input type="checkbox"/> Request Spot instances	
Network	<input type="text" value="vpc-aa8f95d1 (default)"/>	Create new VPC
Subnet	<input type="text" value="No preference (default subnet in any Availability zone)"/>	Create new subnet
Auto-assign Public IP	<input type="text" value="Use subnet setting (Enable)"/>	
Placement group	<input type="checkbox"/> Add instance to placement group	
Capacity Reservation	<input type="text" value="Open"/>	Create new Capacity Reservation
IAM role	<input type="text" value="None"/>	Create new IAM role
Shutdown behavior	<input type="text" value="Stop"/>	
Enable termination protection	<input type="checkbox"/> Protect against accidental termination	
Monitoring	<input type="checkbox"/> Enable CloudWatch detailed monitoring <small>Additional charges apply.</small>	
Tenancy	<input type="text" value="Shared - Run a shared hardware instance"/>	

Cancel Previous Review and Launch Next: Add Storage

7. Ajustar o tamanho do disco se necessário para volumetria do teste e clicar em “**Next: Add Tags**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 4: Add Storage

Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. [Learn more](#) about storage options in Amazon EC2.

Volume Type	Device	Snapshot	Size (GiB)	Volume Type	IOPS	Throughput (MB/s)	Delete on Termination	Encryption
Root	/dev/xvda	snap-0e4c15b8cba3e8ae6	8	General Purpose SSD (gp2)	100 / 3000	N/A	<input checked="" type="checkbox"/>	Not Encrypted

Add New Volume

Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. [Learn more](#) about free usage tier eligibility and usage restrictions.

8. Não utilizaremos Tags. Clicar em “**Next Configure Security Group**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 5: Add Tags

A tag consists of a case-sensitive key-value pair. For example, you could define a tag with key = Name and value = Webserver.
 A copy of a tag can be applied to volumes, instances or both.
 Tags will be applied to all instances and volumes. [Learn more](#) about tagging your Amazon EC2 resources.

Key	(128 characters maximum)	Value	(256 characters maximum)	Instances	Volumes
This resource currently has no tags					
Choose the Add tag button or click to add a Name tag . Make sure your IAM policy includes permissions to create tags.					
Add Tag (Up to 50 tags maximum)					

9. Selecionar o *Security Group* existente que criamos no provisionamento anterior e clicar em “**Review and Launch**”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 6: Configure Security Group

A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. [Learn more](#) about Amazon EC2 security groups.

Assign a security group: Create a new security group
 Select an existing security group

Security Group ID	Name	Description	Actions
sg-0f5dc2bb4c60ec6df	uniritter-aula	Security Group para provisionamento de ambiente de teste	Copy to new
Inbound rules for sg-0f5dc2bb4c60ec6df (Selected security groups: sg-0f5dc2bb4c60ec6df)			
Type	Protocol	Port Range	Source
All traffic	All	All	201.37.162.34/32
SSH	TCP	22	201.37.162.34/32

Cancel **Previous** **Review and Launch**

10. Revisar as configurações e clicar em “Launch”

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 7: Review Instance Launch

Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

▼ AMI Details

Edit AMI



Amazon Linux 2 AMI (HVM), SSD Volume Type - ami-0b69ea66ff7391e80

Free tier eligible

Amazon Linux 2 comes with five years support. It provides Linux kernel 4.14 tuned for optimal performance on Amazon EC2, systemd 219, GCC 7.3, Glibc 2.26, Binutils 2.29.1, and the latest software packages through extras.

Root Device Type: ebs Virtualization type: hvm

▼ Instance Type

Edit instance type

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t2.medium	Variable	2	4	EBS only	-	Low to Moderate

▼ Security Groups

Edit security groups

Security Group ID	Name	Description
sg-0f5dc2bb4c60ec6df	uniritter-aula	Security Group para provisionamento de ambiente de teste

All selected security groups inbound rules

Type	Protocol	Port Range	Source	Description

Cancel Previous Launch

11. Selecionar a chave que criamos no provisionamento anterior e clicar em “Launch”

Select an existing key pair or create a new key pair

X

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about [removing existing key pairs from a public AMI](#).

Choose an existing key pair

Select a key pair

uniritter

I acknowledge that I have access to the selected private key file (uniritter.pem), and that without this file, I won't be able to log into my instance.

Cancel

Launch Instances

12. Acessar a instância provisionada.

```
sh-3.2# ssh -i uniritter.pem ec2-user@54.198.73.67
The authenticity of host '54.198.73.67 (54.198.73.67)' can't be established.
ECDSA key fingerprint is SHA256:q6iNMnUIxQVsB IbZe98Bgw3dJwDYNSh1a0gSLL5k+Hc.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '54.198.73.67' (ECDSA) to the list of known hosts.
Last login: Thu Oct 17 12:43:57 2019 from 201.37.162.34

  _\   _\  )
  \_ (   /  Amazon Linux 2 AMI
  ___\_\_\_\_|_|
```

https://aws.amazon.com/amazon-linux-2/
15 package(s) needed for security, out of 31 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory
[ec2-user@ip-172-31-46-203 ~]\$

13. Instalar o Java 8

Comando: sudo yum -y install java-1.8.0-openjdk

```
[ec2-user@ip-172-31-46-203 ~]$ sudo yum -y install java-1.8.0-openjdk
Failed to set locale, defaulting to C
Loaded plugins: extras_suggestions, langpacks, priorities, update-motd
Resolving Dependencies
--> Running transaction check
--> Package java-1.8.0-openjdk.x86_64 1:1.8.0.222.b10-0.amzn2.0.1 will be installed
--> Processing Dependency: java-1.8.0-openjdk-headless(x86-64) = 1:1.8.0.222.b10-0.amzn2.0.1 for package: 1:java-1
--> Processing Dependency: xorg-x11-fonts-Type1 for package: 1:java-1.8.0-openjdk-1.8.0.222.b10-0.amzn2.0.1.x86_64
--> Processing Dependency: libpng15.so.15(PNG15_0)(64bit) for package: 1:java-1.8.0-openjdk-1.8.0.222.b10-0.amzn2.0.1.x86_64
Processing Dependencies for package libpng15.so.15(PNG15_0)(64bit) for package: 1:java-1.8.0-openjdk-1.8.0.222.b10-0.amzn2.0.1.x86_64
```

14. Realizar download do Kafka

Comando: wget http://ftp.unicamp.br/pub/apache/kafka/2.3.0/kafka_2.12-2.3.0.tgz

```
[ec2-user@ip-172-31-46-203 ~]$ wget http://ftp.unicamp.br/pub/apache/kafka/2.3.0/kafka_2.12-2.3.0.tgz
--2019-10-17 12:47:39--  http://ftp.unicamp.br/pub/apache/kafka/2.3.0/kafka_2.12-2.3.0.tgz
Resolving ftp.unicamp.br (ftp.unicamp.br)... 143.106.10.149
Connecting to ftp.unicamp.br (ftp.unicamp.br)|143.106.10.149|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 57215197 (55M) [application/x-gzip]
Saving to: 'kafka_2.12-2.3.0.tgz'

100%[=====] 2019-10-17 12:47:45 (9.64 MB/s) - 'kafka_2.12-2.3.0.tgz' saved [57215197/57215197]

[ec2-user@ip-172-31-46-203 ~]$
```

15. Descompactar o Kafka

Comando: tar -xzf kafka_2.12-2.3.0.tgz

```
[ec2-user@ip-172-31-46-203 ~]$ tar -xzf kafka_2.12-2.3.0.tgz
[ec2-user@ip-172-31-46-203 ~]$ ls -l
total 55876
drwxr-xr-x 6 ec2-user ec2-user 89 Jun 19 20:44 kafka_2.12-2.3.0
-rw-rw-r-- 1 ec2-user ec2-user 57215197 Jun 25 00:38 kafka_2.12-2.3.0.tgz
[ec2-user@ip-172-31-46-203 ~]$
```

16. Iniciar o Zookeeper

Configurar variável de ambiente para 50% da memoria provisionada no servidor.

Comando: vi .bashrc

Incluir o comando: export KAFKA_HEAP_OPTS="-Xmx2048M -Xms2048M"

```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc
fi
export KAFKA_HEAP_OPTS="-Xmx2048M -Xms2048M"
# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

# User specific aliases and functions
~
~
```

Executar o comando: source .bashrc

Acessar o diretório do Kafka:

```
[ec2-user@ip-172-31-46-203 ~]$ cd kafka_2.12-2.3.0/  
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$  
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$ pwd  
/home/ec2-user/kafka_2.12-2.3.0  
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$ █
```

Executar o comando: nohup bin/zookeeper-server-start.sh config/zookeeper.properties > ~/zookeeper-logs &

```
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$ nohup bin/zookeeper-server-start.sh config/zookeeper.properties > ~/zookeeper-logs &  
[1] 3936  
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$ nohup: ignoring input and redirecting stderr to stdout  
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$ █
```

17. Iniciar o Kafka

Comando: nohup bin/kafka-server-start.sh config/server.properties > ~/kafka-logs &

```
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$ nohup bin/kafka-server-start.sh config/server.properties > ~/kafka-logs &  
[2] 4250  
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$ nohup: ignoring input and redirecting stderr to stdout  
[ec2-user@ip-172-31-46-203 kafka_2.12-2.3.0]$ █
```

Criar tópico de teste:

Comando: bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 13 --topic tst-topic

```
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 13 --topic tst-topic  
Created topic tst-topic.  
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ █
```

Listar tópico:

Comando: bin/kafka-topics.sh --list --zookeeper localhost:2181

```
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ bin/kafka-topics.sh --list --zookeeper localhost:2181  
tst-topic  
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ █
```

Postar mensagem no tópico de teste:

Comando: bin/kafka-console-producer.sh --broker-list localhost:9092 --topic tst-topic

```
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ bin/kafka-console-producer.sh --broker-list localhost:9092 --topic tst-topic  
>mensagem-teste  
>^C[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$  
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ █
```

Verificar mensagem postada anteriormente:

Comando: bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic tst-topic --from-beginning

```
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic tst-topic --from-beginning  
mensagem-teste  
^CProcessed a total of 1 messages  
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ █
```

Para Kafka e Zookeeper:

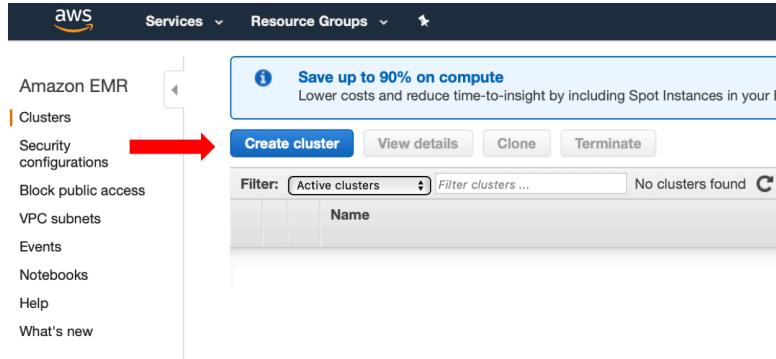
bin/kafka-server-stop.sh

bin/zookeeper-server-stop.sh

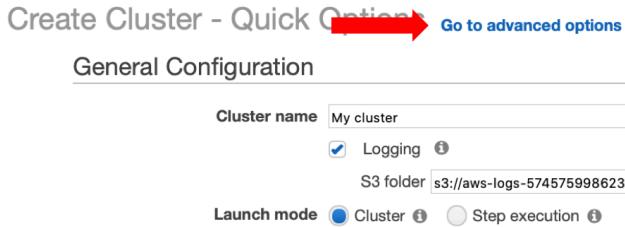
EMR – Elastic MapReduce

Provisionamento de instância EC2 (Elastic Computer Cloud):

1. Realizar *login* na console AWS;
2. Acessar *Services > EMR*
3. Clicar em “Create Cluster”



4. Clicar em “Go to advanced options”



5. Selecionar os componentes conforme abaixo e clicar em “Next”.

The screenshot shows the "Create Cluster - Advanced Options" page. On the left, there's a sidebar with "Step 1: Software and Steps" (selected), "Step 2: Hardware", "Step 3: General Cluster Settings", and "Step 4: Security". The main area is titled "Software Configuration" and shows a "Release" dropdown set to "emr-5.27.0". Underneath, there are two columns of checkboxes for various software components. The first column includes Hadoop 2.8.5, JupyterHub 1.0.0, Ganglia 3.7.2, Hive 2.3.5, MaxMind 1.4.0, Hue 4.4.0, and Spark 2.4.4. The second column includes Zeppelin 0.8.1, Tez 0.9.2, HBase 1.4.10, Presto 0.224, Sqoop 1.4.7, Phoenix 4.14.2, and HCatalog 2.3.5. To the right of these, there are more checkboxes for components like Livy 0.6.0, Flink 1.8.1, Pig 0.17.0, ZooKeeper 3.4.14, Mahout 0.13.0, Oozie 5.1.0, and TensorFlow 1.14.0. Below the component lists, there are sections for "Multi-master support" (unchecked), "AWS Glue Data Catalog settings (optional)" (checkboxes for "Use for Hive table metadata" and "Use for Spark table metadata" checked), "Edit software settings" (radio buttons for "Enter configuration" (selected) and "Load JSON from S3"), and a text input field containing "classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]". At the bottom, there's a "Add steps (optional)" section with a "Step type" dropdown set to "Select a step", a "Configure" button, and a checkbox for "Auto-terminate cluster after the last step is completed". At the very bottom right are "Cancel" and "Next" buttons.

6. Selecionar os componentes conforme abaixo e clicar em “Next”.

- *** Recomenda-se utilizar Spot instances para estes testes devido ao custo.
- *** Configurações de Network e EC2 Subnet conforme sua conta.
- *** tempo aproximado para iniciar o cluster é de 15 minutos.

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

| Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Hardware Configuration ⓘ

If you need more than 20 EC2 instances, see this topic ⓘ.

Instance group configuration Uniform instance groups
Specify a single instance type and purchasing option for each node type.

Instance fleets
Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#) ⓘ

Network Create a VPC ⓘ ⓘ

EC2 Subnet ⓘ ⓘ

Root device EBS volume size 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](#) ⓘ

Node type	Instance type	Instance count	Purchasing option	Auto Scaling
Master	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB ⓘ ⓘ Add configuration settings ⓘ	1 Instances	<input type="radio"/> On-demand ⓘ <input checked="" type="radio"/> Spot ⓘ Use on-demand as max price ⓘ	Not available for Master ⓘ
Core	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB ⓘ ⓘ Add configuration settings ⓘ	2 Instances	<input type="radio"/> On-demand ⓘ <input checked="" type="radio"/> Spot ⓘ Use on-demand as max price ⓘ	Not enabled ⓘ

+ Add task instance group

[Cancel](#) [Previous](#) [Next](#)

7. Informa o nome do cluster.

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

| Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

General Options

Cluster name ⓘ

Logging ⓘ
S3 folder ⓘ

Debugging ⓘ

Termination protection ⓘ

Tags ⓘ

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	<input type="text"/>

Additional Options

EMRFS consistent view ⓘ

Custom AMI ID ⓘ

► Bootstrap Actions

[Cancel](#) [Previous](#) [Next](#)

8. Selecionar a chave e clicar em “Create Cluster”.

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Security Options

EC2 key pair **uniritter**   

Cluster visible to all IAM users in account 

Permissions 

Default Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR_DefaultRole](#) 

EC2 instance profile [EMR_EC2_DefaultRole](#) 

Auto Scaling role [EMR_AutoScaling_DefaultRole](#) 

► Security Configuration

► EC2 security groups

[Cancel](#) [Previous](#) **Create cluster**

9. Aguardar o provisionamento do Cluster.

10. Após finalizar o provisionamento os status devem estar conforme abaixo.

Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Help

What's new

Clone Terminate AWS CLI export

Cluster: uniritter-aula **Running** Running step

Summary Application history Monitoring Hardware Configurations Events Steps Bootstrap actions

Connections: [Enable Web Connection](#) – Hue, Zeppelin, Spark History Server, JupyterHub, Resource Manager ... (View All) 

Master public DNS: ec2-35-175-112-248.compute-1.amazonaws.com [SSH](#)

Tags: -- [View All / Edit](#)

Summary **Configuration details**

ID: j-AZ17ZHPN7GQ9 Release label: emr-5.27.0

Creation date: 2019-10-17 11:13 (UTC-3) Hadoop Amazon 2.8.5

Elapsed time: 8 minutes distribution:

Auto-terminate: No Applications: Hive 2.3.5, Hue 4.4.0, Spark 2.4.4, JupyterHub 1.0.0, Zeppelin 0.8.1, Livy 0.6.0

Termination On [Change](#) protection: Log URI: s3://aws-logs-574575998623-us-east-1/elasticmapreduce/ 

EMRFS consistent view: Disabled

Custom AMI ID: --

Network and hardware **Security and access**

Availability zone: us-east-1f Key name: uniritter

Subnet ID: [subnet-0f8f1400](#) EC2 instance [EMR_EC2_DefaultRole](#) profile:

Master: [Running](#) 1 m5.xlarge EMR role: [EMR_DefaultRole](#)

Spot (max on-demand)

Core: [Running](#) 2 m5.xlarge Auto Scaling role: [EMR_AutoScaling_DefaultRole](#)

Spot (max on-demand)

Task: -- Security groups for [sg-0cda8cf792c0c07af](#) Master: (ElasticMapReduce-master)

Core & Task: (ElasticMapReduce-slave)

Security groups for [sg-080aa2415ae7ec0c7](#)

Configuração do ambiente para ingestão de dados ao Data Lake

1. Acessar Security Group uniritter-aula e liberar acesso as portas/protocolos na rede privada. Clicar no botão “Edit”.

The screenshot shows the AWS EC2 Dashboard with the 'Create Security Group' button and 'Actions' dropdown. On the left, there's a sidebar with various EC2-related options like Instances, Launch Templates, and AMIs. The main area shows a search bar and a table for security groups. One row is selected, labeled 'UNIRITTER'. Below this, a detailed view of the 'Security Group: sg-0f5dc2bb4c60ec6df' is shown. The 'Inbound' tab is selected, and a red arrow points to the 'Edit' button. Another red arrow points to the 'Outbound' tab, which is currently disabled. The table below lists two rules: one for 'All traffic' on port 0-65535 from 'Custom' source 201.37.162.34/32, and another for 'SSH' on port 22 from 'Custom' source 201.37.162.34/32.

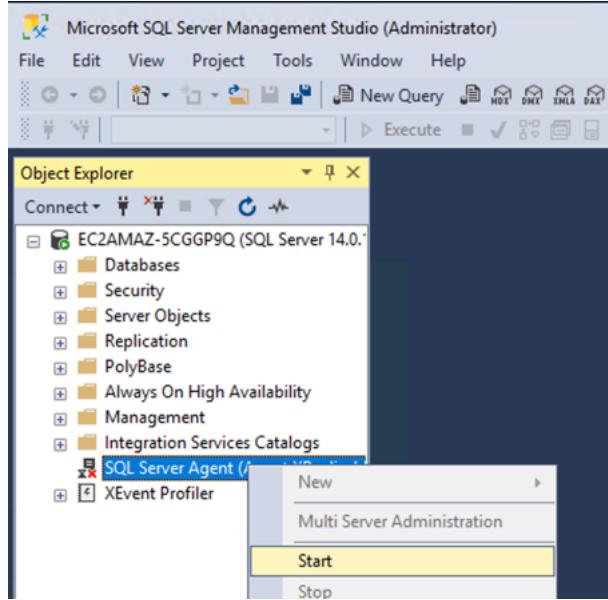
2. Incluir a sua rede privada e clicar em “Save”.

This screenshot shows the 'Edit inbound rules' dialog. It has tabs for Type, Protocol, Port Range, Source, and Description. There are three existing rules: one for 'All traffic' on port 0-65535 from 'Custom' source 201.37.162.34/32 (description: 'e.g. SSH for Admin Desktop'), one for 'SSH' on port 22 from 'Custom' source 201.37.162.34/32 (description: 'e.g. SSH for Admin Desktop'), and one for 'All traffic' on port 0-65535 from 'Custom' source 172.31.0.0/16 (description: 'rede privada'). A new rule is being added with the same parameters. A note at the bottom states: 'NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.' At the bottom right are 'Cancel' and 'Save' buttons.

3. Acessar o SQL Server

This screenshot shows the Microsoft SQL Server Management Studio (Administrator) interface. The title bar says 'Microsoft SQL Server Management Studio (Administrator)'. The 'Object Explorer' pane on the left shows a tree structure for the server 'EC2AMAZ-5CGGP9Q (SQL Server 14.0)'. The branches include Databases, Security, Server Objects, Replication, PolyBase, Always On High Availability, Management, Integration Services Catalogs, SQL Server Agent (Agent XPs disabled), and XEvent Profiler.

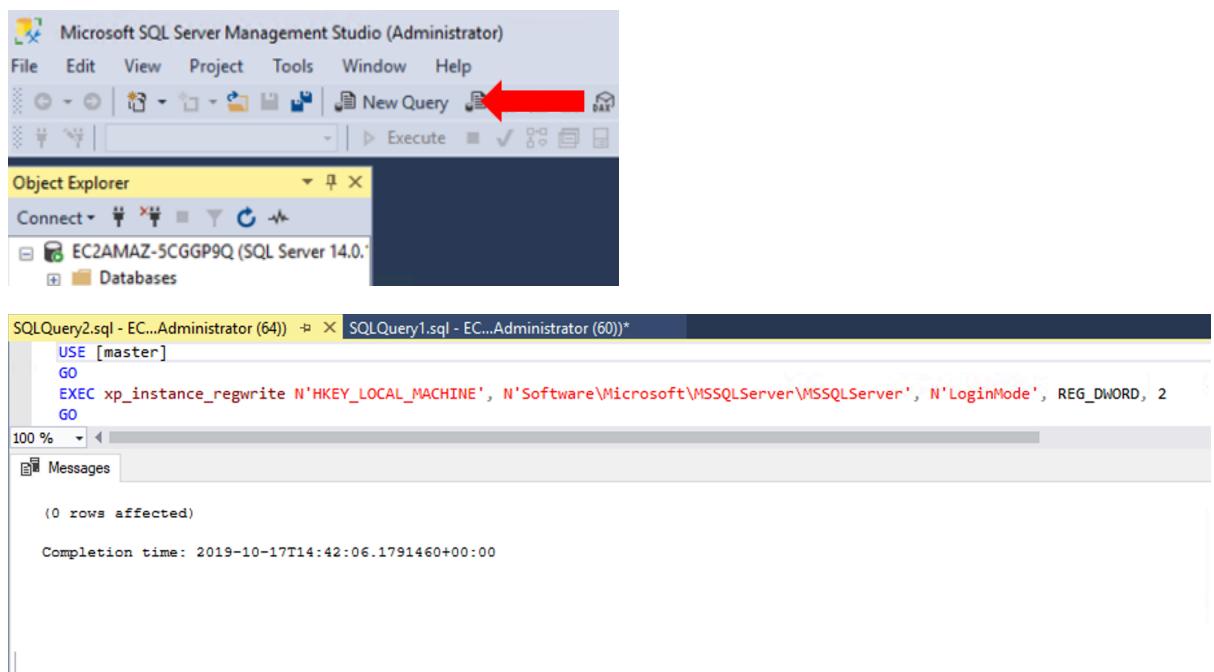
4. Iniciar o SQL Server Agent



5. Selecionar “New Query” e executar o comando abaixo para alteração da instância para modo de autenticação mista (Windows e SQL).

Comando:

```
USE [master]
GO
EXEC xp_instance_regwrite N'HKEY_LOCAL_MACHINE',
N'Software\Microsoft\MSSQLServer\MSSQLServer', N'LoginMode', REG_DWORD, 2
GO
```



6. Criar usuário para os testes.

Comando:

```
USE [master]
GO
CREATE LOGIN [uniritter] WITH PASSWORD=N'uniritter123',
DEFAULT_DATABASE=[master], CHECK_EXPIRATION=OFF, CHECK_POLICY=OFF
GO
ALTER SERVER ROLE [sysadmin] ADD MEMBER [uniritter]
GO
```

The screenshot shows the SQL Server Management Studio interface with three tabs at the top: 'SQLQuery3.sql - EC...Administrator (52)', 'SQLQuery2.sql - EC...Administrator (64)', and 'SQLQuery1.sql - EC...Administrator (60)*'. The 'SQLQuery3.sql' tab contains the T-SQL script provided above. The 'Messages' tab below shows the output: 'Commands completed successfully.' and 'Completion time: 2019-10-17T14:46:47.1489579+00:00'. The status bar at the bottom indicates '100 %' completion.

7. Criar database.

Comando:

```
create database dbUniRitter
go
```

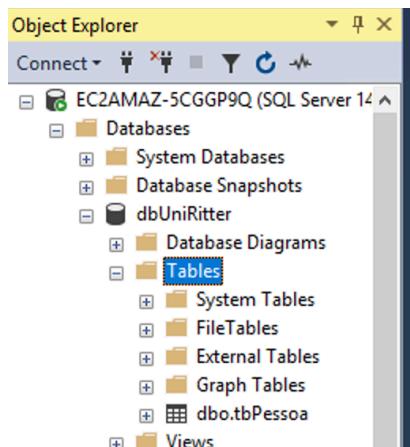
The screenshot shows the SQL Server Management Studio interface with two tabs at the top: 'SQLQuery3.sql - EC...Administrator (52)*' and 'SQLQuery2.sql - EC...Administrator'. The 'SQLQuery3.sql' tab contains the 'create database' command. The 'Messages' tab below shows the output: 'Commands completed successfully.' and 'Completion time: 2019-10-17T14:50:48.1852304+00:00'. The status bar at the bottom indicates '100 %' completion.

8. Criar tabela tbPessoa

Comando:

```
use dbUniRitter
go
create table tbPessoa (
    idPessoa int primary key identity
    ,nmPessoa varchar(150) not null
    ,tpGenero char(1) not null
    ,dtCadastro datetime default getdate()
    ,dtModificacao datetime default getdate()
)
```

```
go
```



Object Explorer

EC2AMAZ-5CGGP9Q (SQL Server 14)

- Databases
 - System Databases
 - Database Snapshots
 - dbUniRitter
 - Database Diagrams
 - Tables
 - System Tables
 - FileTables
 - External Tables
 - Graph Tables
 - dbo.tbPessoa
 - Views

SQLQuery3.sql - EC...Administrator (52)*

```
use dbUniRitter
go
create table tbPessoa (
    idPessoa int primary key identity
    ,nmPessoa varchar(150) not null
    ,tpGenero char(1) not null
    ,dtCadastro datetime default getdate()
    ,dtModificacao datetime default getdate()
)
go
```

100 %

Messages

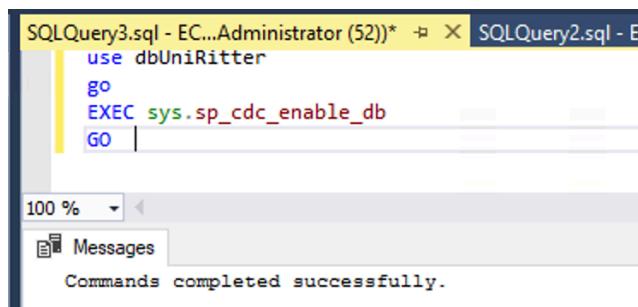
Commands completed successfully.

9. Habilitar CDC – Change Data Capture

Comandos:

```
use dbUniRitter
go
EXEC sys.sp_cdc_enable_db
GO

use dbUniRitter
go
EXEC sys.sp_cdc_enable_table
@source_schema = N'dbo',
@source_name   = N'tbPessoa',
@role_name     = null,
@supports_net_changes = 1
GO
```



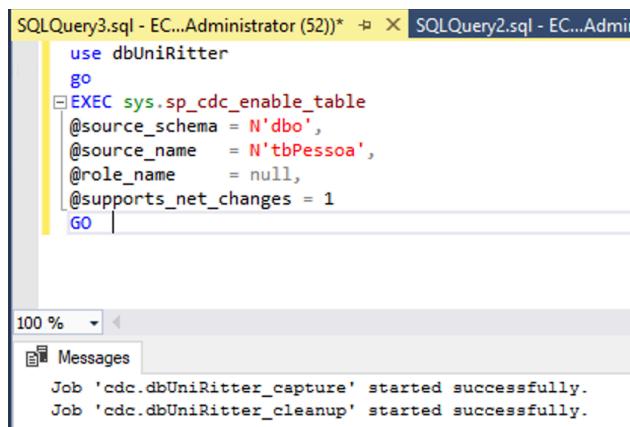
SQLQuery3.sql - EC...Administrator (52)*

```
use dbUniRitter
go
EXEC sys.sp_cdc_enable_db
GO
```

100 %

Messages

Commands completed successfully.



SQLQuery3.sql - EC...Administrator (52)*

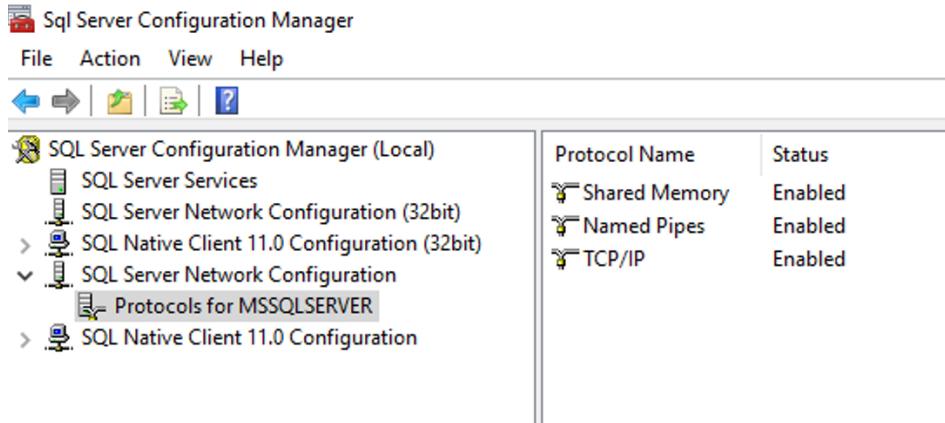
```
use dbUniRitter
go
EXEC sys.sp_cdc_enable_table
@source_schema = N'dbo',
@source_name   = N'tbPessoa',
@role_name     = null,
@supports_net_changes = 1
GO
```

100 %

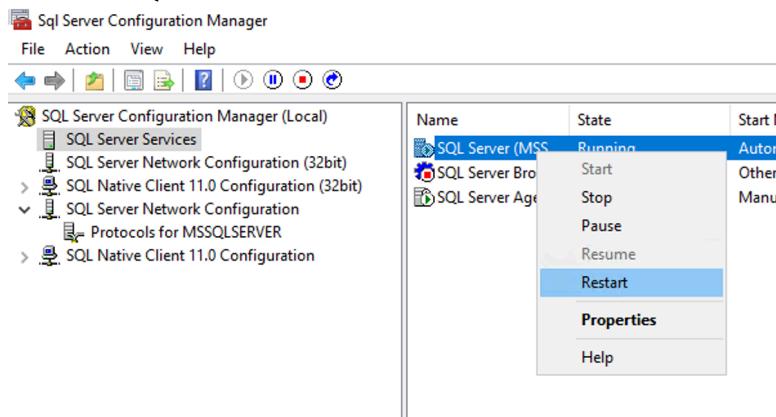
Messages

Job 'cdc.dbUniRitter_capture' started successfully.
Job 'cdc.dbUniRitter_cleanup' started successfully.

10. Acessar o SQL Configuration Manager e habilitar os protocolos conforme abaixo.



11. Reiniciar o SQL Server.



12. Acessar o servidor do Kafka por SSH.

```
ssh -i uniritter.pem ec2-user@<ip do servidor>
```

13. Acessar o diretório do Kafka.

Comando:

```
cd kafka_2.12-2.3.0/
```

14. Criar tópico para receber as mensagens da tabela de pessoa.

Comando:

```
bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 13 --topic kf-tbpessoa
```

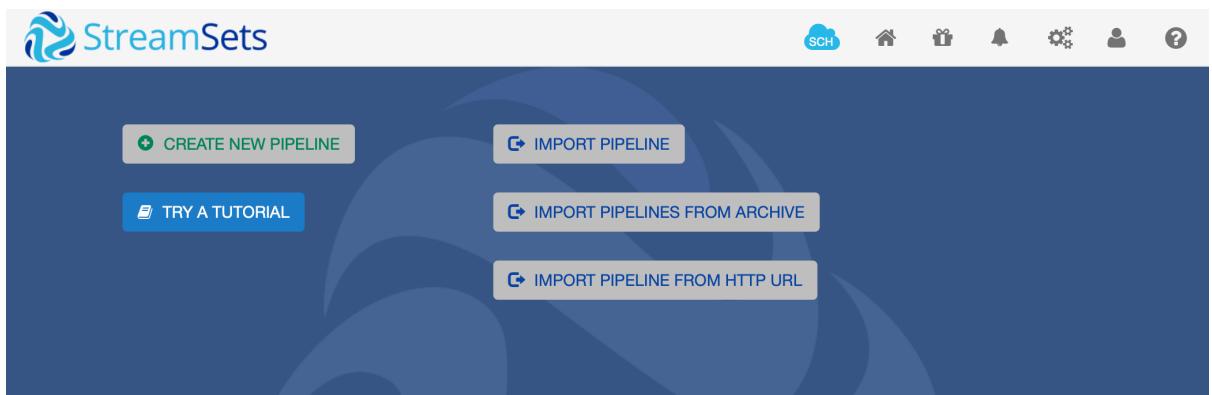
```
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 13 --topic kf-tbpessoa
Created topic kf-tbpessoa.
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$
```

15. Acessar o StreamSets

URL: <http://<ip do servidor>:18630>

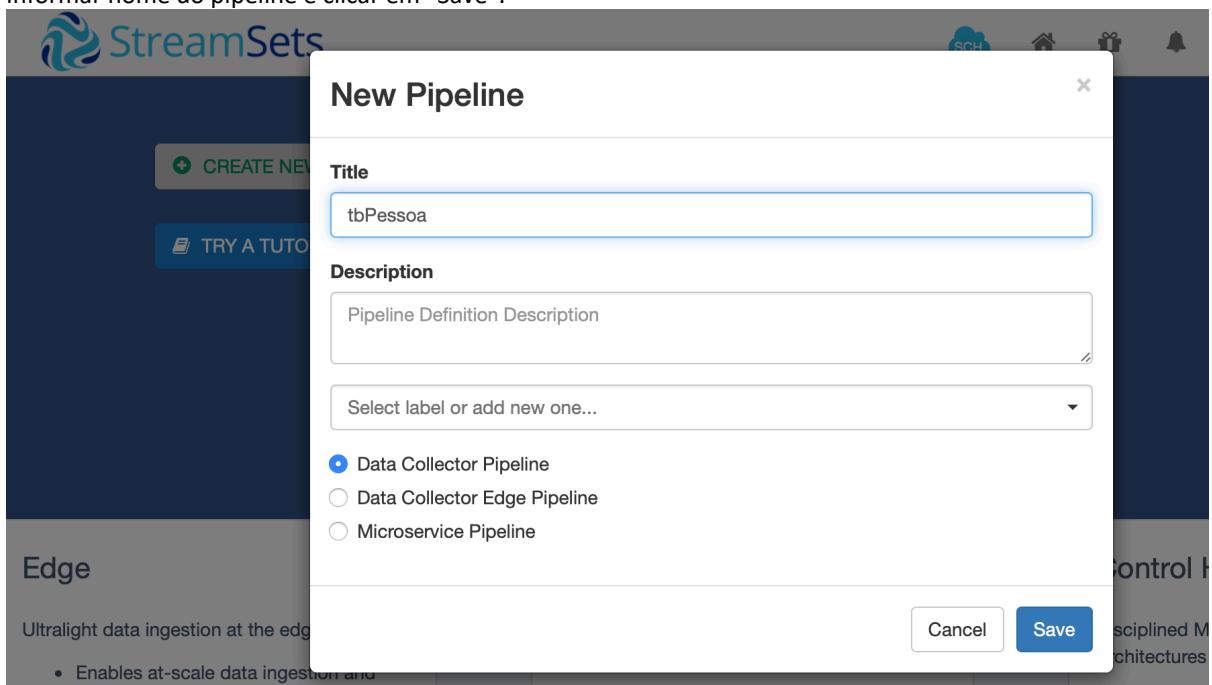
Usuário: admin

Senha: admin



16. Clicar em Create New Pipeline.

17. Informar nome do pipeline e clicar em “Save”.



18. Pipeline criado.

The screenshot shows the StreamSets Data Pipeline interface. At the top, there's a toolbar with various icons like 'SCH', 'Home', 'Import', 'Export', etc. Below the toolbar, the title bar says 'Pipelines / tbPessoa'. A message 'Origin missing' with a dropdown 'Select Origin...' is displayed. On the right, there's a 'Don't show again.' checkbox. The main workspace is a grid where a red warning icon is visible. On the left, there's a sidebar with tabs: 'Info' (selected), 'Configuration' (highlighted in blue), 'Rules', and 'History'. The 'Configuration' tab contains fields for 'Pipeline ID' (tbPessoa7c32a8d4-7d27-4b20-9c22-5141471c6603), 'Title' (tbPessoa), 'Description', 'Labels', and 'Execution Mode' (set to 'Standalone'). A 'Display a menu' button is at the bottom of the sidebar.

19. Selecionar SQL Server CDC Client.

The screenshot shows the StreamSets Data Pipeline interface. The title bar says 'Pipelines / tbPessoa'. A message 'SQL Server CDC Client 1 has open stream' is shown above two dropdown menus: 'Select Processor to connect...' and 'Or Select Destination to connect..'. To the right, there's a sidebar titled 'Origins' with a dropdown set to 'cdc'. It lists 'Oracle CDC Client' and 'PostgreSQL CDC Client'. At the bottom, there's a section for 'Add/Remove Stages' with a plus sign icon. A red warning icon is visible on the grid.

20. Configurar JDBC.

Ex. JDBC - jdbc:sqlserver://172.31.46.211:1433;databaseName=dbUniRitter;

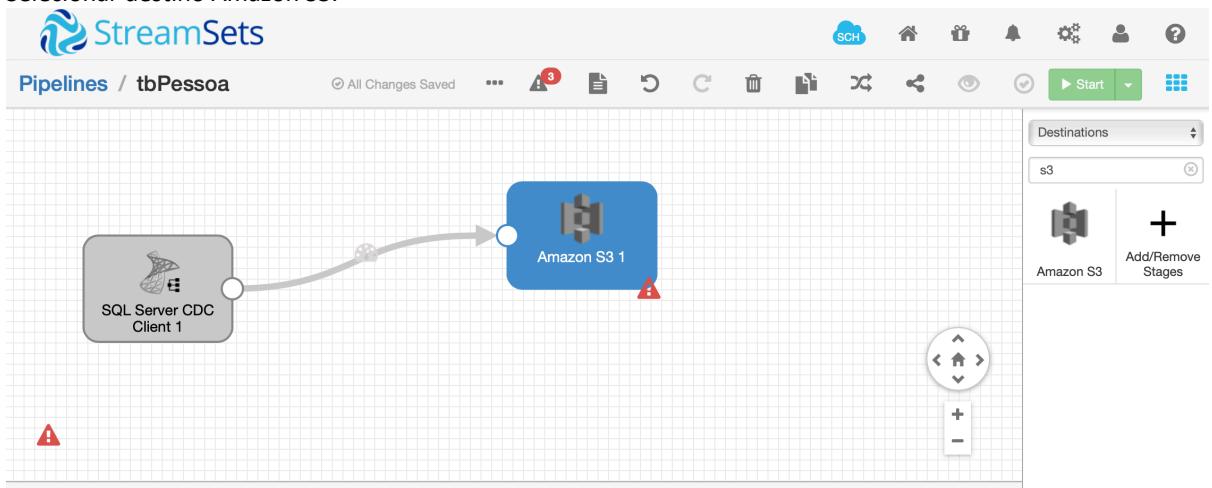
The screenshot shows the StreamSets Data Pipeline interface. At the top, there is a toolbar with various icons and a status message: "SQL Server CDC Client 1 has open stream". Below the toolbar, there is a central workspace with a blue rounded rectangle containing a small icon and the text "SQL Server CDC Client 1". A red warning icon is positioned below it. On the right side of the workspace, there is a circular navigation control with arrows and symbols. At the bottom of the workspace, there is a dropdown menu labeled "SQL Server CDC Client 1".

The main configuration area has tabs: Info, General, JDBC, CDC, Credentials, Legacy, and Advanced. The "JDBC" tab is selected. Under the "Configuration" section, the "JDBC Connection String" field contains the value "jdbc:sqlserver://172.31.46.211:1433;databaseName=dbUniRitter;". The "Use Credentials" checkbox is checked. Below that, "Queries Per Second" is set to 10 and "Number of Threads" is set to 1. There is also a checkbox for "Use Direct Table Query" which is unchecked.

21. Informar credenciais para o JDBC.

This screenshot is identical to the previous one, showing the StreamSets Data Pipeline interface for configuring a "SQL Server CDC Client" processor. The configuration screen is identical, with the "Credentials" tab selected and the "Username" field set to "uniritter" and the "Password" field set to "uniritter123".

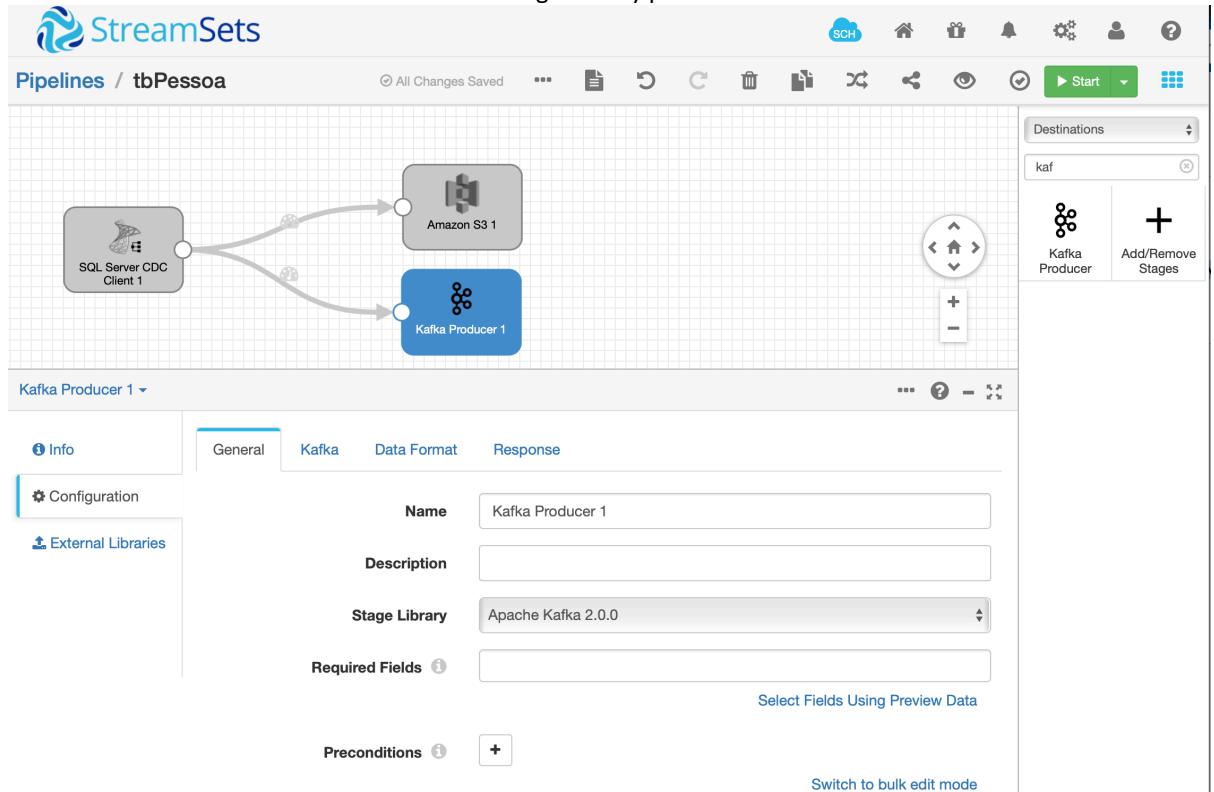
22. Selecionar destino Amazon S3.



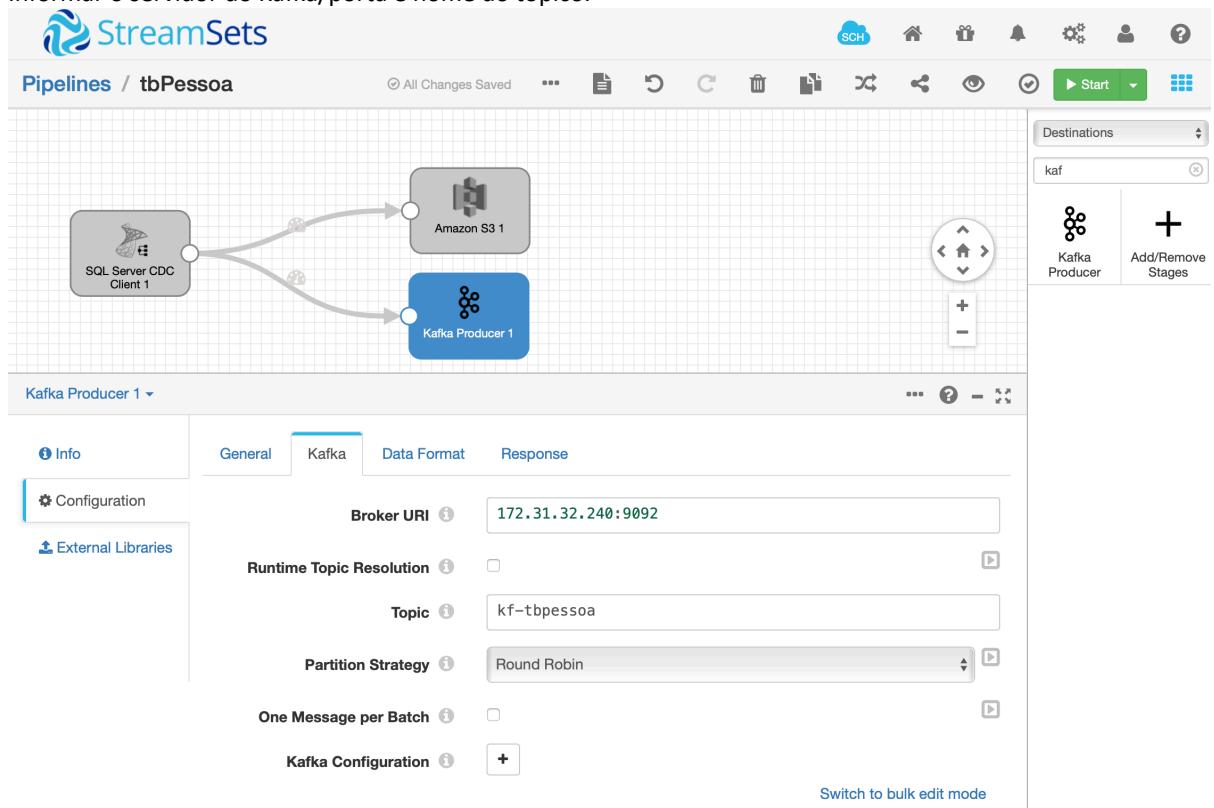
23. Configurar credenciais de acesso ao S3, região e bucket.

This screenshot provides a detailed view of the StreamSets Pipeline Editor's configuration panel for the "Amazon S3 1" stage. The configuration tab "Amazon S3" is selected. The panel includes fields for "Access Key ID" (with a placeholder value), "Secret Access Key" (with a placeholder value), "Region" (set to "US East (N. Virginia) us-east-1"), and "Bucket" (set to "datalake01-raw/tbpessoal"). On the left, a sidebar lists "Info", "General", "Amazon S3" (which is active), "SSE", "Advanced", and "Data Format". Other sections like "Configuration", "External Libraries", and "Generated Events" are also visible.

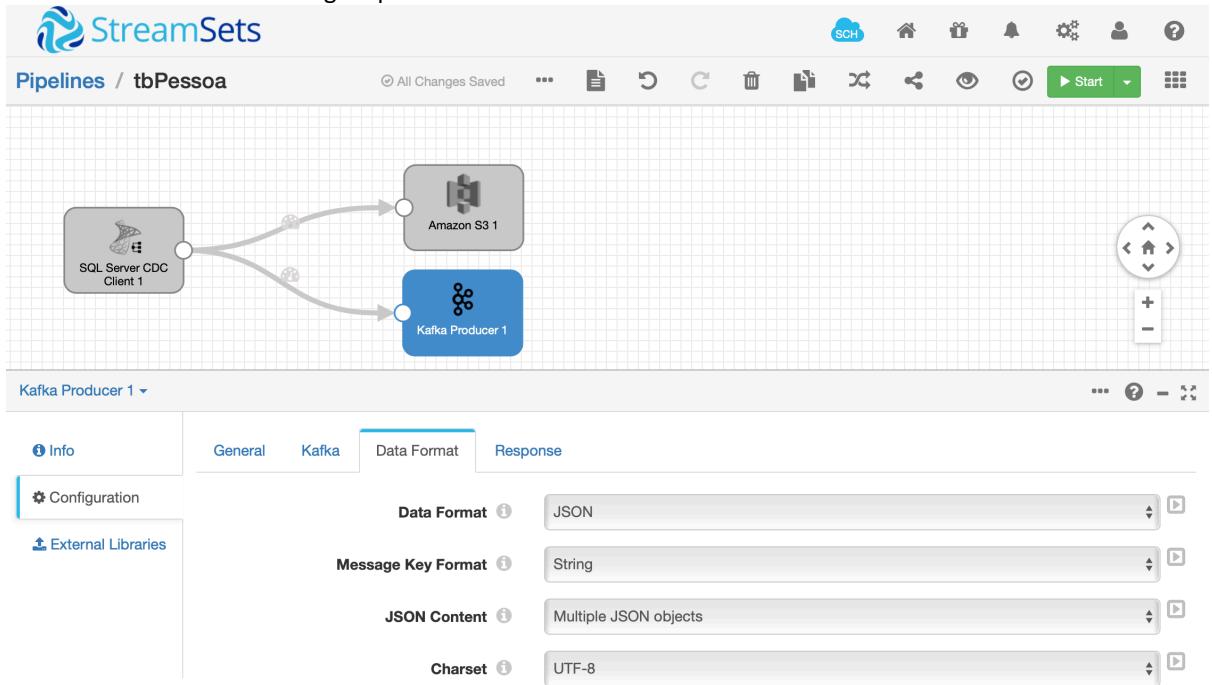
24. Selecionar destino Kafka Producer e alterar o Stage Library para versão abaixo.



25. Informar o servidor do Kafka/porta e nome do tópico.



26. Selecionar formato de mensagem para o Kafka.



27. Iniciar o pipeline.

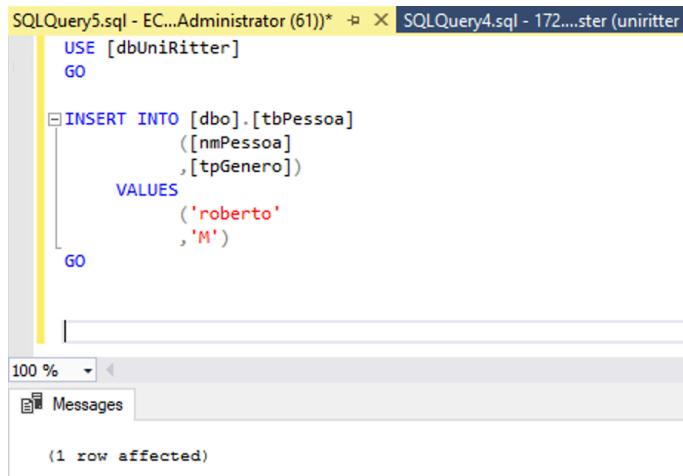


28. Realizar teste de fluxo de ingestão inserindo registro na tabela de pessoa no SQL Server.

Comando:

```
USE [dbUniRitter]
GO
```

```
INSERT INTO [dbo].[tbPessoa]
([nmPessoa]
,[tpGenero])
VALUES
('roberto'
,'M')
GO
```

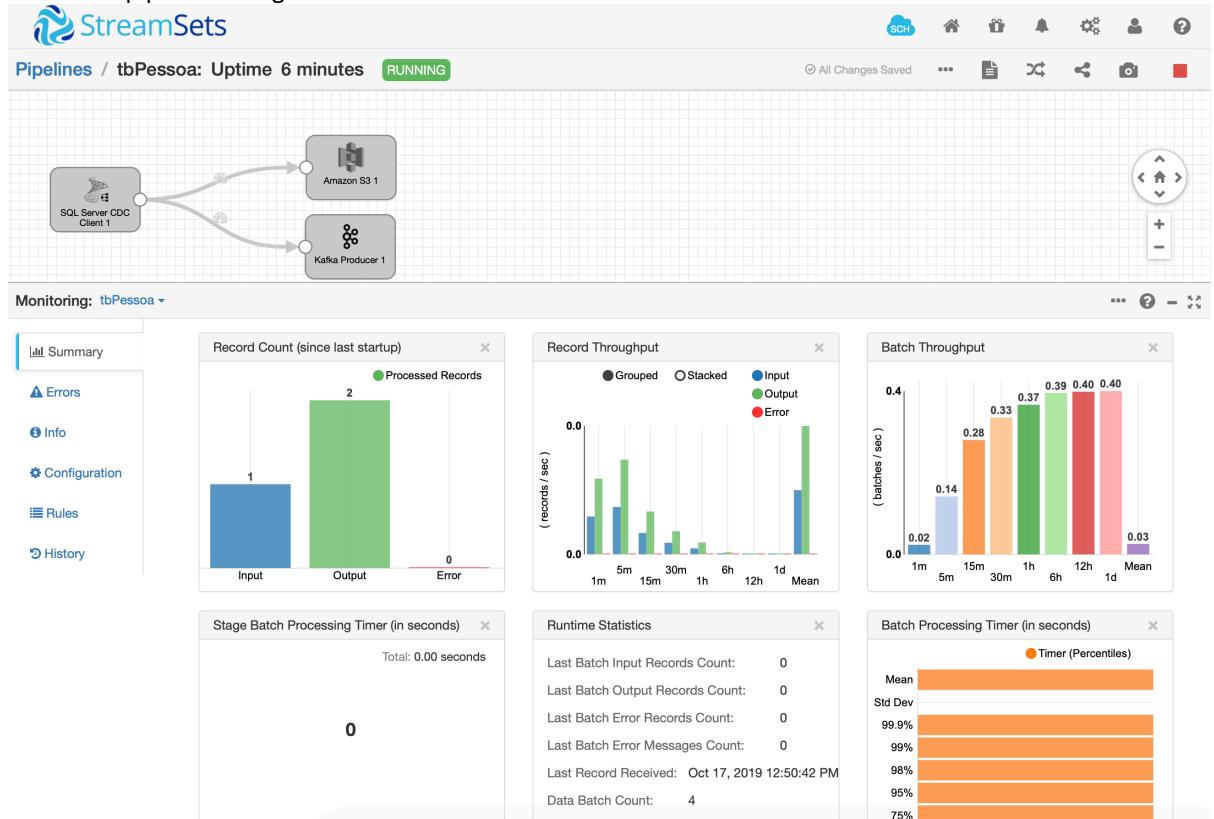


```
SQLQuery5.sql - EC...Administrator (61)*  X  SQLQuery4.sql - 172...ster (uniritter)
USE [dbUniRitter]
GO

INSERT INTO [dbo].[tbPessoa]
([nmPessoa]
,[tpGenero])
VALUES
('roberto'
,'M')
GO
```

(1 row affected)

29. Verificar o pipeline de ingestão.



Disciplina: Processamento de Grandes Volumes de Dados

Professor: Roberto Galvão

StreamSets

Pipelines / tbPessoa: Uptime 6 minutes RUNNING

Monitoring: SQL Server CDC Client 1 ▾

```
graph LR; A[SQL Server CDC Client 1] --> B(( )); B --> C[Amazon S3 1]; B --> D[Kafka Producer 1]
```

Table Metrics for Thread - 0.0

Status:	WAITING_FOR_RATE_LIM
Current Table:	cdc.dbo_tbPessoa_CT
Tables Owned:	1
Thread Name:	Table Jdbc Runner - 0
Last Rate Limit Wait (sec)	0.095829

Records Processed (since last startup)

Record Count (since last startup)

Record Throughput

Batch Processing Timer (in seconds)

Records Per Batch Histogram (5 minutes...)

Summary

- Errors
- Info
- Configuration
- External Libraries
- Generated Events

StreamSets

Pipelines / tbPessoa: Uptime 7 minutes RUNNING

Monitoring: Amazon S3 1 ▾

```
graph LR; A[SQL Server CDC Client 1] --> B(( )); B --> C[Amazon S3 1]; B --> D[Kafka Producer 1]
```

Records Processed (since last startup)

Record Count (since last startup)

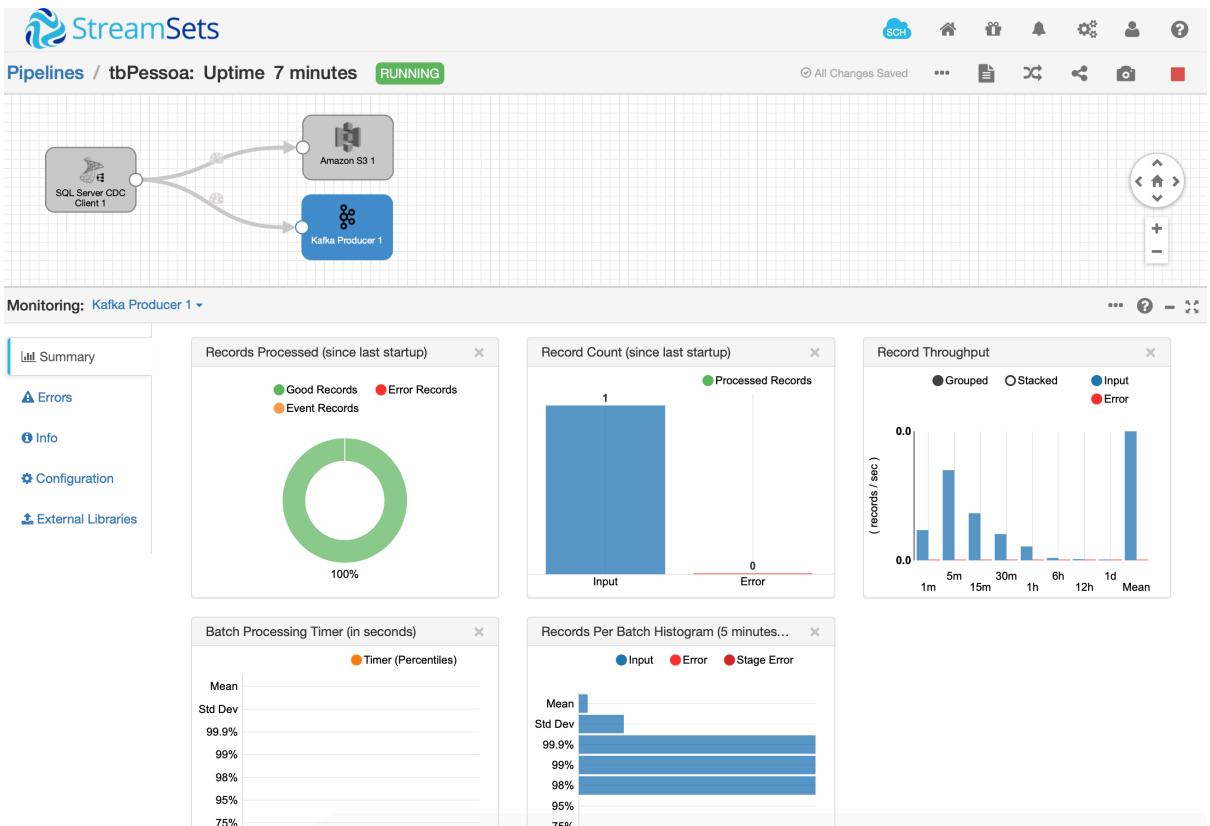
Record Throughput

Batch Processing Timer (in seconds)

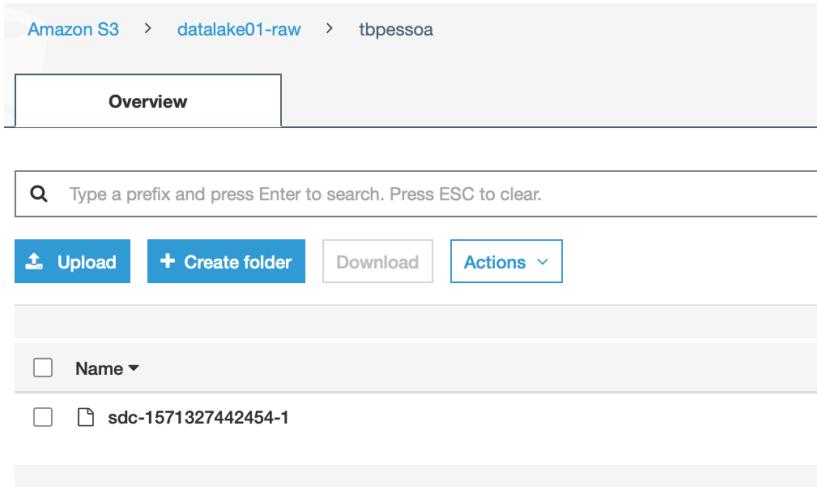
Records Per Batch Histogram (5 minutes...)

Summary

- Errors
- Info
- Configuration
- External Libraries
- Generated Events



30. Verificar o recebimento no bucket S3.



The screenshot shows the Amazon S3 console for the "datalake01-raw" bucket. It displays an "Overview" section with a search bar and a list of objects. One object is visible: "sdc-1571327442454-1".

31. Verificar o recebimento no tópico do Kafka.

```
[ec2-user@ip-172-31-32-240 ~]$ cd kafka_2.12-2.3.0/  
[ec2-user@ip-172-31-32-240 kafka_2.12-2.3.0]$ bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic kf-tbpessoa --from-beginning  
{"idPessoa":1,"nmPessoa":"roberto","tpGenero":"M","dtCadastro":1571327438900,"dtModificacao":1571327438900}
```

32. Criar tabela no AWS Athena para realizar consulta nos dados.

Acessar Services > Athena

Comando:

```
CREATE EXTERNAL TABLE datalake01_tbpessoa (
    idPessoa int
    ,nmPessoa string
    ,tpGenero string
    ,dtCadastro string
    ,dtModificacao string
)
ROW FORMAT serde 'org.apache.hive.hcatalog.data.JsonSerDe'
LOCATION 's3://datalake01-raw/tbpessoa/';
```

The screenshot shows the AWS Athena Query Editor interface. On the left, the Database dropdown is set to 'default' and the Tables section shows one entry: 'datalake01_tbpessoa'. Below it, the Views section is empty. On the right, the main area contains a code editor with the following SQL query:

```
1 CREATE EXTERNAL TABLE datalake01_tbpessoa (
2     idPessoa int
3     ,nmPessoa string
4     ,tpGenero string
5     ,dtCadastro string
6     ,dtModificacao string
7 )
8 ROW FORMAT serde 'org.apache.hive.hcatalog.data.JsonSerDe'
9 LOCATION 's3://datalake01-raw/tbpessoa/';
```

Below the code editor, there are several buttons: 'Run query' (highlighted in blue), 'Save as', 'Create', and 'Format query'. A status message indicates '(Run time: 0.78 seconds, Data scanned: 0 KB)'. At the bottom, a results panel displays the message 'Query successful.'.

33. Realizar consultar nos arquivos inseridos no datalake.

The screenshot shows the AWS Athena Query Editor interface. On the left, the 'Database' section is set to 'default' with a table named 'datalake01_tbpessoa'. The main area displays a query window with three tabs: 'New query 1', 'New query 2', and 'New query 3' (which is currently active). The query code is:

```
1 SELECT idpessoa, nm pessoa, tpgenero
2 FROM "default"."datalake01_tbpessoa" limit 10;
```

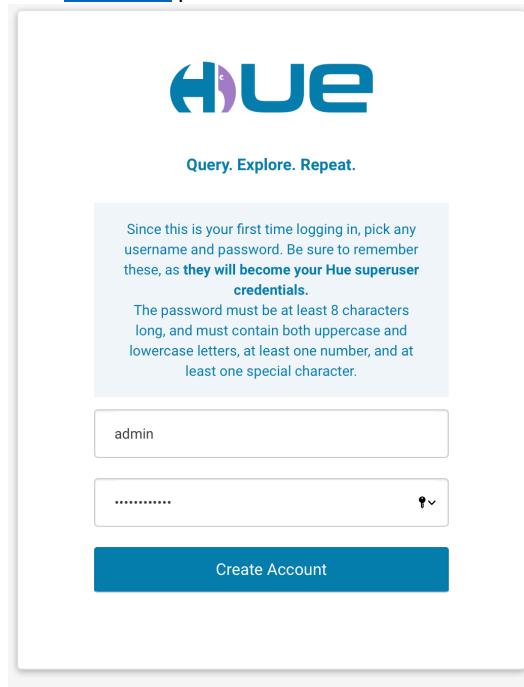
Below the query window, there are buttons for 'Run query', 'Save as', 'Create', and 'Format query'. A status message indicates a run time of 1.49 seconds and data scanned of 0.1 KB. The results section shows a single row of data:

	idpessoa	nm pessoa	tpgenero
1	1	roberto	M

34. Realizar acesso através do self-service (HUE).

No primeiro acesso é necessário definir usuário e senha.

Link: <http://<ip publico>:8888>



35. Visualização da estrutura de dados.

The screenshot shows the Hue Table Browser interface. At the top, there's a navigation bar with icons for Home, Help, and Logout. Below it is a toolbar with icons for Tables, Queries, and Refresh. The main area shows a breadcrumb path: Databases > default > datalake01_tbpessoa. On the left, there's a sidebar with a back arrow, a folder icon, and a search bar labeled "Filter...". The main content area displays the table schema:

Column (5)	Type	Description	Sample
i idpessoa	int	Add a description...	1
i nmpessoa	string	Add a description...	roberto
i tpgenero	string	Add a description...	M
i dtcadastro	string	Add a description...	1571327438900
i dtmodificacao	string	Add a description...	1571327438900

36. Consulta através do HIVE.

The screenshot shows the Apache Hive interface. At the top, there are tabs for 'Hive' and 'SQL', and buttons for 'Add a name...', 'Add a description...', and three icons. Below the tabs, it says '3.22s Database default Type text ?'. The query editor contains the following code:

```
1 select idpessoa, nmPessoa, tpGenero
2 from datalake01_tbPessoa
```

A modal window shows the execution log:

```
Select idpessoa, nmPessoa, tpGenero
from datalake01_tbPessoa
INFO : Completed executing command(queryId=hive_20191017173942_b19e13e5-8cc2-4857-ba3f-7c79
0fb1e380); Time taken: 0.001 seconds
INFO : OK
```

Below the log, the 'Results (1)' tab is selected, showing the following table:

	idpessoa	nmPessoa	tpGenero
1	1	roberto	M

37. Consulta através do SparkSQL.

The screenshot shows the Apache SparkSQL interface. At the top, there are tabs for 'SQL' and 'PySpark', and buttons for 'Add a name...' and 'Add a description...'. Below the tabs, it says '2.11s Database default Type text ?'. The query editor contains the same code as the Hive screenshot:

```
1 select idpessoa, nmPessoa, tpGenero
2 from datalake01_tbPessoa
```

A modal window shows the execution log:

```
INFO : CONCURRENCY_MODE IS DISABLED, NOT CREATING A LOCK MANAGER
INFO : Executing command(queryId=hive_20191017174436_94a0950c-e2d8-4ba1-8c3b-54f27ae471c2): select idpessoa, nmPessoa, tpGenero
from datalake01_tbPessoa
INFO : Completed executing command(queryId=hive_20191017174436_94a0950c-e2d8-4ba1-8c3b-54f27ae471c2); Time taken: 0.002 seconds
INFO : OK
```

Below the log, the 'Results (1)' tab is selected, showing the same table as the Hive screenshot:

	idpessoa	nmPessoa	tpGenero
1	1	roberto	M

38. Notebook PySpark.

The screenshot shows the PySpark Notebook interface. At the top, there are tabs for 'PySpark' and 'SQL', and buttons for 'Add a name...' and 'Add a description...'. Below the tabs, it says '10.58s ?'. The code editor contains:

```
1 df = spark.read.json("s3://datalake01-raw/tbPessoa")
2 df.show()
```

A modal window shows the execution log:

```
at py4j.commands.Command$1.execute(Command$1.java:74)
at py4j.GatewayConnection.run(GatewayConnection.java:238)
at java.lang.Thread.run(Thread.java:748)
```

On the right, it says 'application_1571333279633_0001'. Below the log, the 'Results (1)' tab is selected, showing the following table:

	dtCadastro	dtModificacao	idPessoa	nmPessoa	tpGenero
1	1571327438900	1571327438900	1	roberto	M

39. Notebook Scala.

The screenshot shows a Scala notebook interface. At the top, there are tabs for 'Scala' and 'PySpark', and buttons for 'Add a name...', 'Add a description...', and three icons. The status bar indicates '11.71s ?'. Below the tabs is a code editor with the following Scala code:1 val df = spark.read.json("s3://datalake01-raw/tbpessoa/")
2 df.show()
3 |A large scrollable area displays YARN logs and diagnostics. The logs show executor registration and a new executor being registered. The diagnostics section shows a single row of data from the DataFrame:| idPessoa | nmPessoa | tpGenero |
| --- | --- | --- |
| 1571327438900 | roberto | M |

40. Acesso ao contexto do HIVE através do PySpark.

```
from pyspark.sql import HiveContext
hive_context = HiveContext(sc)
df = hive_context.sql("select idpessoa, nm pessoa from default.datalake01_tbpessoa ")
df.show()
```

The screenshot shows a PySpark notebook interface. At the top, there are tabs for 'PySpark' and 'Scala', and buttons for 'Add a name...', 'Add a description...', and three icons. The status bar indicates '19.73s ?'. Below the tabs is a code editor with the same PySpark code as above:1 from pyspark.sql import HiveContext
2 hive_context = HiveContext(sc)
3 df = hive_context.sql("select idpessoa, nm pessoa from default.datalake01_tbpessoa ")
4 df.show()A large scrollable area displays YARN logs and diagnostics. The logs show task starting and ACK_LOCAL. The diagnostics section shows a single row of data from the DataFrame:| idpessoa | nm pessoa |
| --- | --- |
| 1 | roberto |

41. Visualização de bucket S3.

The screenshot shows the AWS S3 File Browser interface. At the top, there's a search bar labeled "Search for file name", an "Actions" dropdown, a "Delete forever" button, and two buttons for "Upload" and "New". Below the header, the path "us-east-1 s3a:// datalake01-raw / tbpessoa" is displayed. The main area is a table with the following data:

	Name	Size	User	Group	Permissions	Date
					drwxrwxrwx	October 17, 2019 10:45 AM
					drwxrwxrwx	
		107 bytes			-rw-rw-rw-	October 17, 2019 08:50 AM

At the bottom left, it says "Show 45 of 1 items". On the right, there's a "Page" indicator showing "1 of 1" and navigation icons.

42. Notebook Zeppelin.

The screenshot shows the Zeppelin Notebook interface with three cells of Scala code:

```
%spark
val df = spark.read.json("s3://datalake01-raw/tbpessoa/")
df.show()
```

```
+-----+-----+-----+
| dtCadastro|dtModificacao|idPessoa|nmPessoa|tpGenero|
+-----+-----+-----+
|1571327438900|1571327438900|    11| roberto|     M|
+-----+-----+-----+
```

df: org.apache.spark.sql.DataFrame = [dtCadastro: bigint, dtModificacao: bigint ... 3 more fields]


```
%spark.pyspark
df = spark.read.json("s3://datalake01-raw/tbpessoa/")
df.show()
```

```
+-----+-----+-----+
| dtCadastro|dtModificacao|idPessoa|nmPessoa|tpGenero|
+-----+-----+-----+
|1571327438900|1571327438900|    11| roberto|     M|
+-----+-----+-----+
```



```
%spark.pyspark
from pyspark.sql import HiveContext
hive_context = HiveContext(sc)
df = hive_context.sql("select idpessoa, nmPessoa from default.datalake01_tbpessoa ")
df.show()
```

```
+-----+
|idpessoanmPessoal|
+-----+
|    11| roberto|
+-----+
```

Each cell has a "SPARK JOBS FINISHED" status indicator at the top right.

Disciplina: Processamento de Grandes Volumes de Dados

Professor: Roberto Galvão

43. Notebook JupyterHub (EMR)

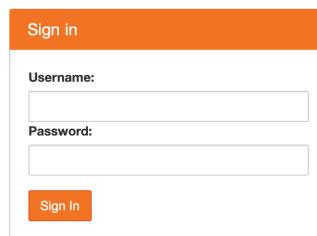
<https://<ip> publico:9443>

Usuário default: jovyan

Senha: jupyter



Jupyter



44. Pyspark

Jupyter PySpark Last Checkpoint: a minute ago (unsaved changes)

Logout Control Panel

```
File Edit View Insert Cell Kernel Widgets Help Trusted PySpark3 O


In [2]: df = spark.read.json("s3://datalake01-raw/tbpessoa/")
df.show()
+-----+-----+-----+
| dtCadastro | dtModificacao | idPessoa | nmPessoa | tpGenero |
+-----+-----+-----+
| 1571327438900 | 1571327438900 | 1 | roberto | M |
+-----+-----+-----+
In [3]: from pyspark.sql import HiveContext
hive_context = HiveContext(sc)
df = hive_context.sql("select idpessoa, nmpessoa from default.datalake01_tbpessoa ")
df.show()
+-----+
| idpessoa | nmpessoa |
+-----+
| 1 | roberto |
+-----+
```

45. Spark – Scala

Jupyter Scala Last Checkpoint: a minute ago (unsaved changes)

Logout Control Panel

```
File Edit View Insert Cell Kernel Widgets Help Trusted Spark O


In [2]: val df = spark.read.json("s3://datalake01-raw/tbpessoa/")
df.show()
df: org.apache.spark.sql.DataFrame = [dtCadastro: bigint, dtModificacao: bigint ... 3 more fields]
+-----+-----+-----+
| dtCadastro | dtModificacao | idPessoa | nmPessoa | tpGenero |
+-----+-----+-----+
| 1571327438900 | 1571327438900 | 1 | roberto | M |
+-----+-----+-----+
```