



Inovação com dados em nuvem

Ingestão de Dados do
Twitter
Guia para Laboratório
Hands-On

Alexandre Alves Andrade
Junho de 2021



Este trabalho está licenciado sob uma Licença Creative Commons Atribuição-Compartilha Igual 4.0 Internacional. Para ver uma cópia desta licença, visite <http://creativecommons.org/licenses/by-sa/4.0/>.

Guia para Laboratório *Hands-On*

Introdução	4
Lab. 1 – Acessando seu ambiente	5
Lab. 2 – Criando e configurando seu ambiente de ingestão de dados	8
Criação de conta no Twitter	8
Criação de VM na Oracle Cloud	9
Lab. 4 – Criando a ingestão com NiFi.	16

Introdução

Um dos motores da adoção de uma cultura data-driven está em justamente ingerir, catalogar e distribuir dados diversos ao maior número possível de pessoas dentro da companhia, permitindo que decisões sejam embasadas em informações precisas e não somente percepções.

E a ingestão a um repositório centralizado é o motor para a democratização dos dados em todas as camadas da empresa.

Neste laboratório vamos explorar um dos aspectos dessa ingestão, o streaming de dados. Em nosso exemplo iremos carregar dados a partir do Twitter para um assunto em específico, que poderá ser utilizado posteriormente para mapear nossos clientes, entender a reação sobre nossos concorrentes ou mesmo com as pessoas tem visto os nossos produtos.

Para a execução deste *hands on* vamos precisar de uma conta de desenvolvedor no Twitter (você poderá criar a sua a partir deste [link](#)) e uma máquina virtual na Oracle Cloud Infrastructure. No demais utilizaremos serviços open source, como NiFi, Metabase e MySQL.

É importante que os conceitos fundamentais desses recursos estejam claros para uma boa experiência em nossa nuvem.

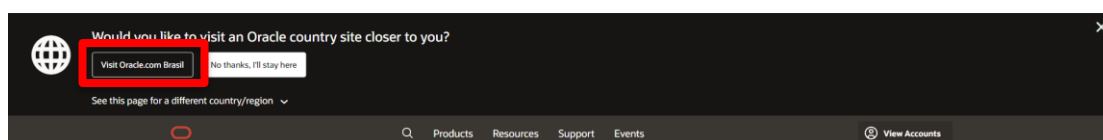
Lab. 1 – Acessando seu ambiente

Objetivos

- Acessar o console da Oracle Cloud
- Conhecer os serviços de infraestrutura e plataforma
- Familiarizar-se com o ambiente

Nesta seção você aprenderá mais sobre o acesso inicial ao ambiente.

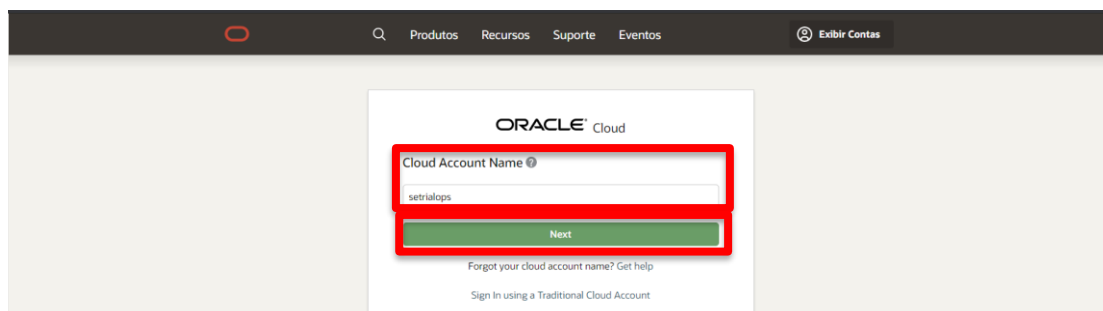
Vá para oracle.com. Você pode alterar o idioma dessa página antes do acesso ao ambiente:



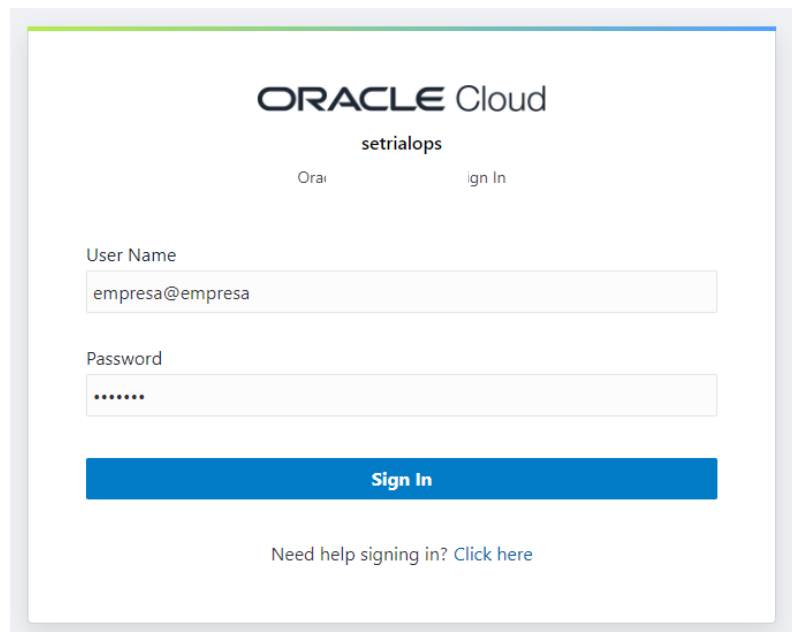
No site já em português, clique em Exibir Contas e depois em Faça Login na Nuvem:



O login deve ser feito com o “Cloud Account Name”, onde somente é necessário informar o **nome da conta** (definido no momento de solicitação). Caso você ainda não tenha a sua conta prossiga para a opção **Inscrever-se no Modo Gratuito na Nuvem**.



A tela de login é apresentada. O usuário administrador é identificado pelo e-mail utilizado no cadastro do ambiente.

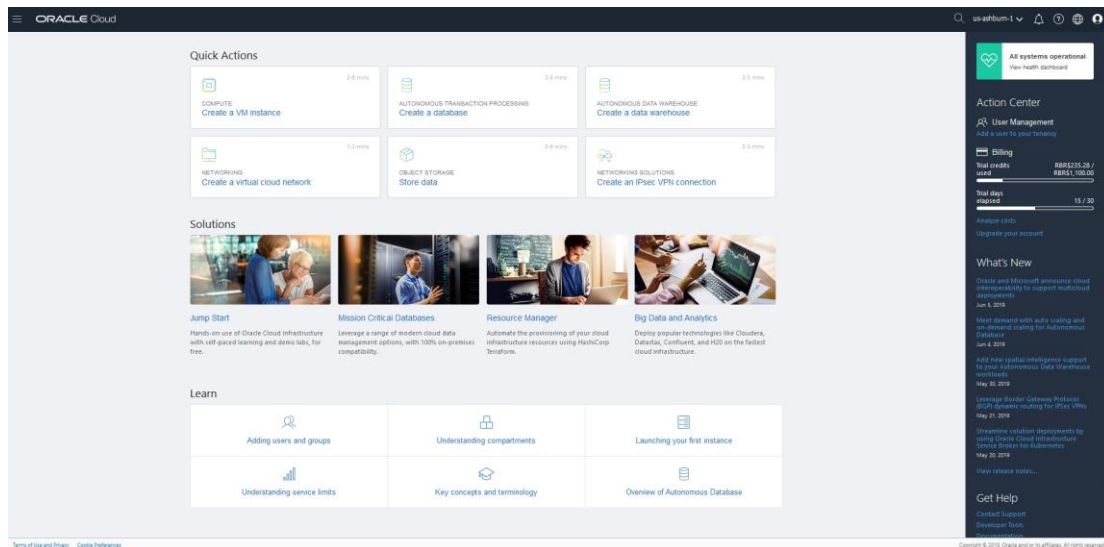


The screenshot displays the Oracle Cloud login interface. At the top, the 'ORACLE Cloud' logo is centered. Below it, the text 'setrialops' is visible, followed by 'Ora' and 'ign In'. The main form contains a 'User Name' field with the text 'empresa@empresa' and a 'Password' field with masked characters '.....'. A blue 'Sign In' button is positioned below the password field. At the bottom, there is a link that says 'Need help signing in? Click here'.

A tela principal do seu ambiente é apresentada. Nela, você consegue ver algumas ações rápidas para a criação de alguns recursos, alguns artigos de soluções dentro da nuvem da Oracle que possa ajudar, a parte de Learn que leva para a documentação, que é muito bem detalhada.

Na barra superior tem-se o menu que lista todas as abas da console da nuvem e os seguintes componentes:

- Lupa: busca no ambiente OCI tanto de serviços quanto de documentação;
- Região: indicação sobre a região ativa. São 30+ regiões no mundo, sendo 2 regiões no Brasil (São Paulo e Vinhedo);
- Sino: anúncios relacionados à sua nuvem;
- Interrogação: atalho para ajuda sobre a página que você estiver navegando assim como para interação com o suporte da Oracle;
- Chat: assistente virtual para a Oracle Cloud que poderá te ajudar com dúvidas de maneira geral e te levar para um atendimento com um agente da Oracle, caso necessário;
- Mundo: configuração da língua da sua interface do Oracle Cloud;
- Perfil: atalho para informações e configurações de sua nuvem (tenancy) e seu usuário.



Feito o acesso, a primeira coisa que será feita é configurar um Compartimento. Não é recomendado criar nenhum recurso no compartimento root, portanto, para a criação de um compartimento basta ir ao menu (canto superior esquerdo), Identidade e Segurança -> Identidade -> Compartimentos. Para mais detalhes acesse este link:

<https://docs.cloud.oracle.com/iaas/Content/Identity/Tasks/managingcompartments.htm>

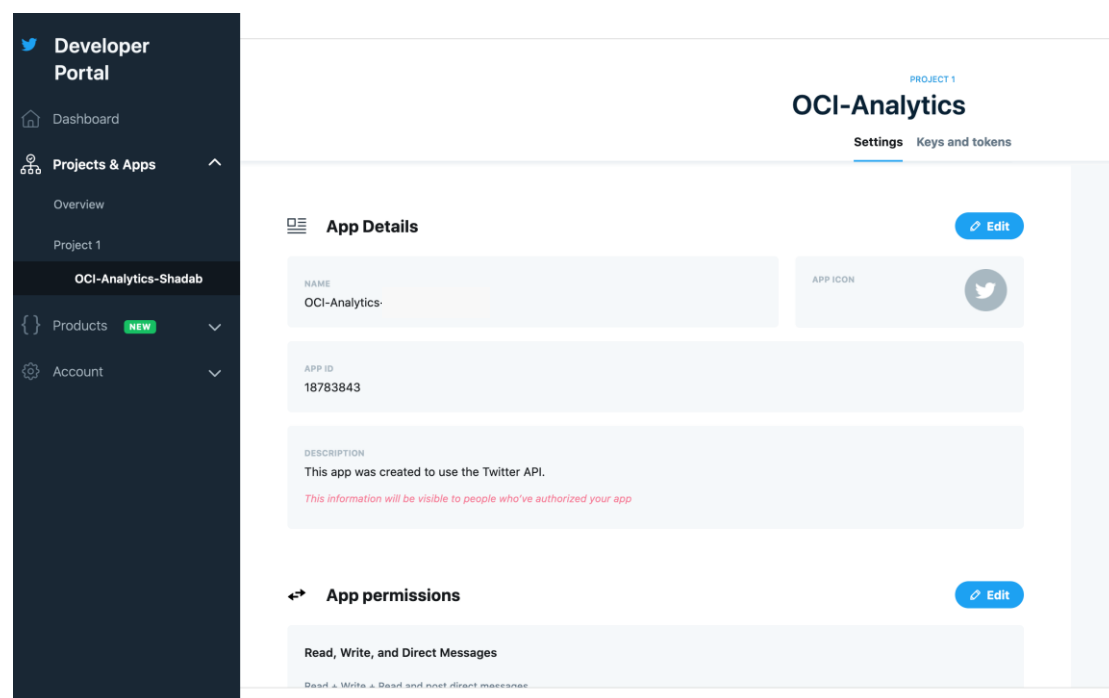
Lab. 2 – Criando e configurando seu ambiente de ingestão de dados

Objetivos:

- Configuração de sua conta de desenvolvedor no Twitter;
- Criação e configuração de Máquina Virtual;
- Acesso ao NiFi

Criação de conta no Twitter

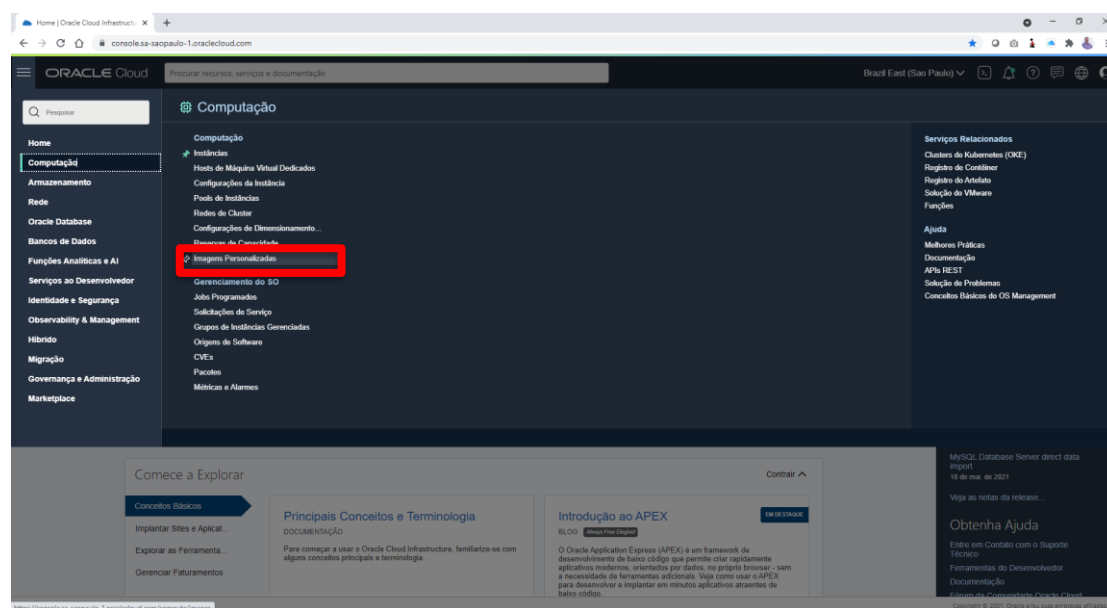
Acesse a página do Twitter para criação de [conta de desenvolvedor](#) e após a liberação, que pode variar de minutos a dias, crie uma aplicação, onde serão gerados os consumer key, consumer secret, access token e access secret. Guarde essas informações com cuidado e não compartilhe com ninguém. O Twitter só te mostrará essas informações uma vez.



Criação de VM na Oracle Cloud

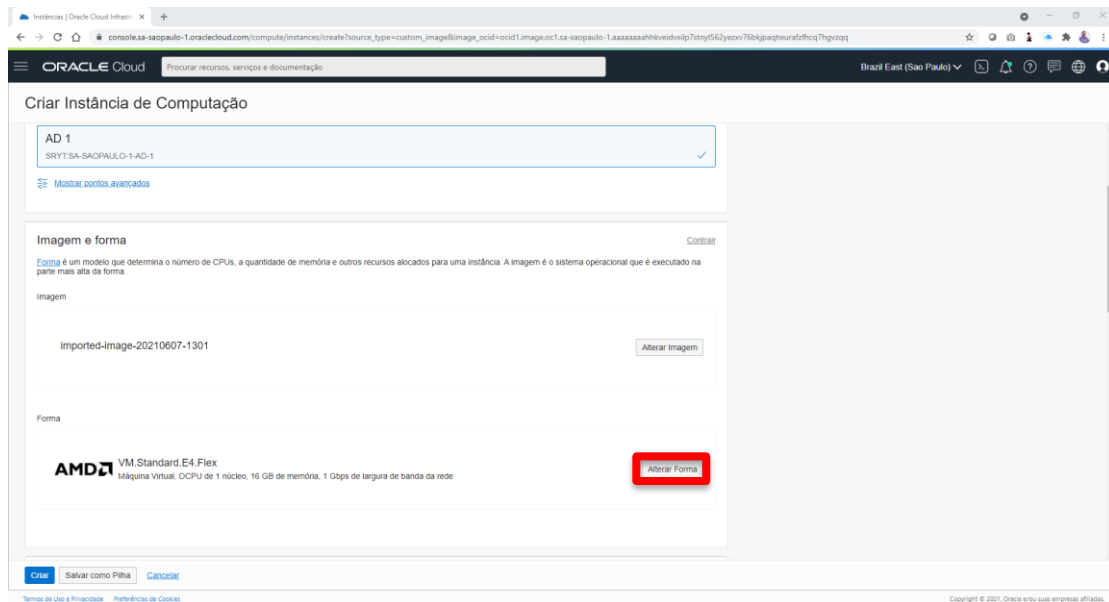
A sua VM será criada a partir de uma imagem customizada criada pelo [Alexandre Alves](#) a partir do modelo disponibilizado pelo [Prof. Fábio Jardim](#).

Acesse o menu no canto superior esquerdo, clique em Computação -> Instâncias -> Imagens Personalizadas.

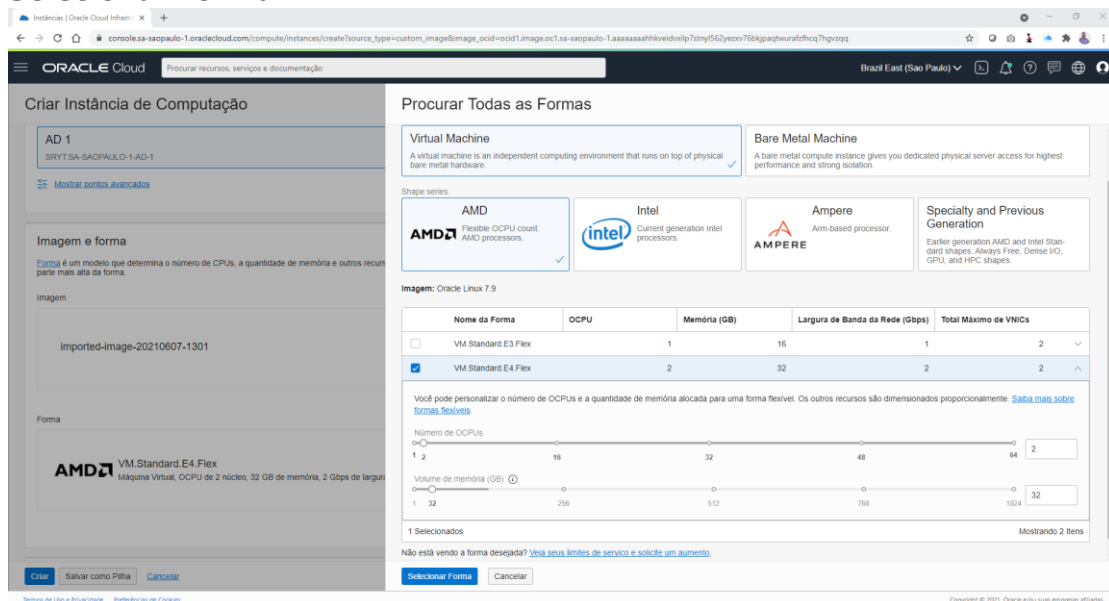


Escolha o compartimento criado anteriormente e clique no botão **Importar Imagem**. Dê um nome para a sua imagem, mantenha o sistema operacional com Linux e selecione a opção **Importar de um URL do Object Storage**. Na caixa **URL do Serviço Object Storage** coloque o seguinte link, para que seja copiado para sua cloud (tenancy) a imagem da VM que foi preparada: <https://objectstorage.sa-saopaulo-1.oraclecloud.com/n/oraclemetodista/b/VMs/o/VM%20Big%20Data%20TDC%20Connections%202021>

Em tipo de imagem escolha a opção **OCI** e clique no botão **Importar Imagem**.



Na janela que se abre confirme a opção de shape AMD com a forma VM.Standard.E4.Flex. Selecione 2 OCPUs e 32 GB de RAM. Clique em **Selecionar Forma**.



IMPORTANTE: essa VM irá consumir seus créditos (se você estiver no período de *trial* irá consumir de seus créditos para avaliação), portanto, após a realização do laboratório recomenda-se que ela seja desligada, consumindo assim apenas os recursos de armazenamento (disco de boot e imagem personalizada). Caso você queira cessar por completo todo o consumo, a VM deverá ser excluída, assim como a imagem personalizada. **Realizada a exclusão todos os dados desta VM serão perdidos de maneira irreversível.**

Selecione a rede virtual (VCN) onde sua VM será criada. Caso não tenha nenhuma criada clique na opção **Criar uma nova rede virtual na nuvem**, dê

um nome para a rede e para a subnet. Crie essa rede no mesmo compartimento de sua VM e mantenha a sugestão do Bloco CIDR.

Oracle Cloud console - Criar Instância de Computação

Rede

Rede é a forma como a instância estabelece conexão com a internet e com outros recursos na Console. Para ter certeza de que você possa estabelecer conexão com sua instância, designe a ela um endereço IP público.

Rede

☐ Selecionar rede virtual na nuvem existente ☒ Criar uma nova rede virtual na nuvem ☐ Digitar o OCID da sub-rede

Nome da nova rede virtual na nuvem: VCN_TDC_Innovation

Criar no compartimento: TDCInnovation

Sub-rede

☐ Selecionar sub-rede existente ☒ Criar uma nova sub-rede pública

Nome da nova sub-rede: Public_Subnet_TDC_Innovation

Criar no compartimento: TDCInnovation

Bloco CIDR: 10.0.0.0/24

Endereço IP Público

☒ Designar um endereço IPv4 público ☐ Não designar um endereço IPv4 público

A designação de um endereço IP público torna essa instância acessível pela internet. Se não tiver certeza de que precisa de um endereço IP público, você poderá sempre designar um posteriormente.

[Mostrar pontos avançados](#)

Criar Salvar como Pilha Cancelar

Para acesso via ssh à VM será necessário um par de chave pública e privada. Caso você não tenha seu par é possível salvar as chaves clicando nos botões correspondentes na opção **Generate a key pair for me**. Caso prefira utilizar as suas chaves escolha a opção **Fazer upload dos arquivos de chave pública (.pub)** ou **Colar chaves públicas**.

Mantenha as opções padrões para o **Volume de inicialização**. Clique em **Criar**.

Oracle Cloud console - Criar Instância de Computação

Adicionar chaves SSH

Generate an [SSH key pair](#) to connect to the instance using SSH, or upload a public key that you already have.

☒ Generate a key pair for me ☐ Fazer upload dos arquivos de chave pública (.pub) ☐ Colar chaves públicas ☐ Nenhuma chave SSH

Faça download da chave privada para que você possa estabelecer conexão com a instância usando SSH. Essa mensagem não será exibida novamente.

[Salvar Chave Privada](#) [Salvar Chave Pública](#)

Volume de inicialização

O [volume de inicialização](#) é um dispositivo que pode ser desanexado e que contém a imagem usada para inicializar a instância de computação.

☐ Especifique um tamanho do volume de inicialização personalizado
[Exemplo do volume](#) varia conforme o tamanho do volume. Tamanho do volume de inicialização padrão: 40 GB

☒ Usar criptografia em trânsito
[Criptografia em trânsito](#) entre a instância, o volume de inicialização e os volumes em blocos.

☐ Criptografe este volume com uma chave gerenciada por você
Por padrão, a Oracle gerencia as chaves que criptografam esse volume, mas você pode escolher uma chave de um vault ao qual você tenha acesso ou criar um controle maior sobre o ciclo de vida da chave e como ela é utilizada. [Saiba mais sobre como renovar suas chaves de criptografia](#)

[Mostrar pontos avançados](#)

Criar Salvar como Pilha Cancelar

Após alguns minutos a sua VM estará disponível. Copie o endereço IP público e utilize seu cliente favorito de SSH para o acesso (Putty, MobaXTerm, diretamente pelo prompt do seu sistema operacional, etc.). Você deverá

utilizar o usuário *opc* e indicar o caminho completo da chave privada que você salvou no passo anterior. Ela substituirá a senha.

The screenshot displays the Oracle Cloud console interface. At the top, the navigation bar shows 'ORACLE Cloud' and a search bar. The main content area is titled 'VM TDC Innovation Big Data' and includes a green status indicator labeled 'EXECUTANDO'. Below this, there are tabs for 'Informações da Instância', 'Oracle Cloud Agent', and 'Tags'. The 'Informações da Instância' tab is active, showing various details:

- Informações Gerais:** Domínio de Disponibilidade: AD-1, Domínio de Falha: FD-3, Região: sa-saopaulo-1, OCID: ...xp2woq, Iniciado: seg., 7 de jun. de 2021 16:48:04 UTC, Compartimento: oraclemetadista (root) (raiz)/SE-ZANI/SE-ALEXANDRE-ANDRADE/TCOTDCInnovation, Tipo de capacidade: Sub demanda.
- Detalhes da Instância:** Rede Virtual na Nuvem: VCN_TDC_Innovation, Reinicialização de Manutenção: -, Imagem: imported-image-20210607-1301, Modo de Início: NATIVE, Serviço de Metadados da Instância: Versões 1 e 2, Migração Ativa: Desativado, Ação de Recuperação de Manutenção: Restaurar instância.
- Configuração de Forma:** Forma: VM Standard E4 Flex, Contagem de OCPUs: 2.
- Acesso da Instância:** Endereço IP Público: 129.159.55.144, Endereço IP Privado: 10.0.0.202, Grupos de Segurança de Rede: Nenhum, Sub-rede: Public_Subnet_TDC_Innovation, Registro de DNS privado: Ativar, Nome do Host: vm-tdc-innovation-big-data, FQDN Interno: vm-tdc-innovation-big-data...
- Opções de Início:** Tipo de Anexação de NIC: VFIO, Volume de Dados Remoto: PARAVIRTUALIZED, Firmware: UEFI_64, Tipo de Volume de Inicialização: PARAVIRTUALIZED, Criptografia em Trânsito: Desativado.

Below the console, a terminal window is open, showing the SSH connection details and the command prompt for the user 'opc' on the host 'opc@vm-tdc-innovation-big-data'.

Essa VM tem um servidor WEB assim como vários outros serviços de Big Data. Para acessá-los vamos precisar abrir as portas que serão utilizadas. Ainda na mesma janela onde a instância foi criada cliente no nome da rede.

Computação » Instâncias » Detalhes da Instância » Solicitações de Serviço

VM TDC Innovation Big Data

[Iniciar](#)
[Interromper](#)
[Reinicializar](#)
[Editar](#)
[Mais Ações](#)

Informações da Instância Oracle Cloud Agent Tags

Informações Gerais

Domínio de Disponibilidade: AD-1
Domínio de Falha: FD-3
Região: sa-saopaulo-1
OCID: ...xp2xoq [Mostrar](#) [Copiar](#)
Iniciado: seg., 7 de jun. de 2021 16:48:04 UTC
Compartimento: oraclemetodista (root) (raiz)/SE-ZANI/SE-ALEXANDRE-ANDRADE/TDC/TDCInnovation
Tipo de capacidade: Sob demanda

Detalhes da Instância

Rede Virtual na Nuvem: **VCN_TDC_Innovation**
Reinicialização de Manutenção: -
Imagem: [imported-image-20210607-1301](#)

Na janela que se abre clique no nome da Sub-rede. Na janela seguinte clique na **Default Security List**.

Rede » Redes Virtuais na Nuvem » Detalhes da Rede Virtual na Nuvem

VCN_TDC_Innovation

[Mover Recurso](#)
[Adicionar Tags](#)
[Encerrar](#)

Informações da VCN Tags

Compartimento: TDCInnovation
Criado: seg., 7 de jun. de 2021 16:47:57 UTC
Bloco CIDR IPv4: 10.0.0.0/16
Bloco CIDR do IPv6: Nenhum Valor

Rede » Redes Virtuais na Nuvem » VCN_TDC_Innovation » Detalhes da Sub-rede

Public_Subnet_TDC_Innovation

[Editar](#)
[Mover Recurso](#)
[Adicionar Tags](#)
[Encerrar](#)

Informações de Sub-rede Tags

OCID: ...wbizla [Mostrar](#) [Copiar](#)
Bloco CIDR IPv4: 10.0.0.0/24
Endereço Mac de Roteador Virtual: 00:00:17:B2:72:9C
Tipo de Sub-rede: Regional

Recursos

Sub-redes em TDCInnovation Compartimento

[Criar Sub-rede](#)

Nome	Estado
Public_Subnet_TDC_Innovation	● Disponível

Logos

Filtros de Tag [adicionar](#) [limpar](#)

nenhum filtro de tag aplicado

Listas de Segurança

[Adicionar Lista de Segurança](#)

Nome	Estado
Default Security List for VCN_TDC_Innovation	● Disponível

Na nova janela clique no botão **Adicionar Regras de Entrada**. Configure com as seguintes opções:

- CIDR de Origem: 0.0.0.0/0
- Intervalo de portas de destino:
80,50070,8080,16010,8081,9000,3000,9090,8889,8888,4040
- Não é necessário alterar as demais opções

Clique em **Adicionar Regras de Entrada**. A nova janela deve aparecer dessa maneira:

Oracle Cloud

Regras de Entrada

Adicionar Regras de Entrada Editar Remover

	Sem Monitoramento de Estado	Origem	Protocolo IP	Intervalo de Portas de Origem	Intervalo de Portas de Destino	Tipo e Código	Permite	Descrição
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	22		TCP tráfego nas portas: 22 SSH Remote Login Protocol	
<input type="checkbox"/>	Não	0.0.0.0/0	ICMP			3, 4	Tráfego ICMP para: 3, 4 Destino Inaceessível: Fragmentação Necessária e foi definido Não Fragmentar	
<input type="checkbox"/>	Não	10.0.0.0/16	ICMP			3	Tráfego ICMP para: 3 Destino Inaceessível	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	80		TCP tráfego nas portas: 80	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	50070		TCP tráfego nas portas: 50070	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	8080		TCP tráfego nas portas: 8080	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	16010		TCP tráfego nas portas: 16010	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	8081		TCP tráfego nas portas: 8081	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	9000		TCP tráfego nas portas: 9000	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	3000		TCP tráfego nas portas: 3000	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	9090		TCP tráfego nas portas: 9090	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	8889		TCP tráfego nas portas: 8889	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	8888		TCP tráfego nas portas: 8888	
<input type="checkbox"/>	Não	0.0.0.0/0	TCP	Todos	4040		TCP tráfego nas portas: 4040	

0 Selecionados

Mostrando 14 itens < 1 de 1

Para testar abra uma nova janela de seu browser e coloque o endereço `http://<ip_da_sua_vm>` onde você substituirá o `<ip_da_sua_vm>` pelo IP copiado anteriormente. Você deverá ver uma janela como essa:

TDC Connections 2021

Inovação com Dados em Nuvem

Divirta-se em nosso ambiente criado especialmente para você, como parte da palestra Cultura Data-Driven da trilha Inovação com Dados em Nuvem.

Este ambiente foi criado pelo [Alexandre Alves](#) a partir do modelo disponibilizado pelo Prof. Fábio Jardim a quem sou muito grato pelo apoio.

- Nifi
- Mongo Express
- HDPFS
- Hbase
- Kafka Manager
- Metabase
- Jupyter Spark
- Hue

Nos acompanhe em nossas redes, inscreva-se em nosso [canal do YouTube](#) e participe de nosso grupo no [LinkedIn](#).

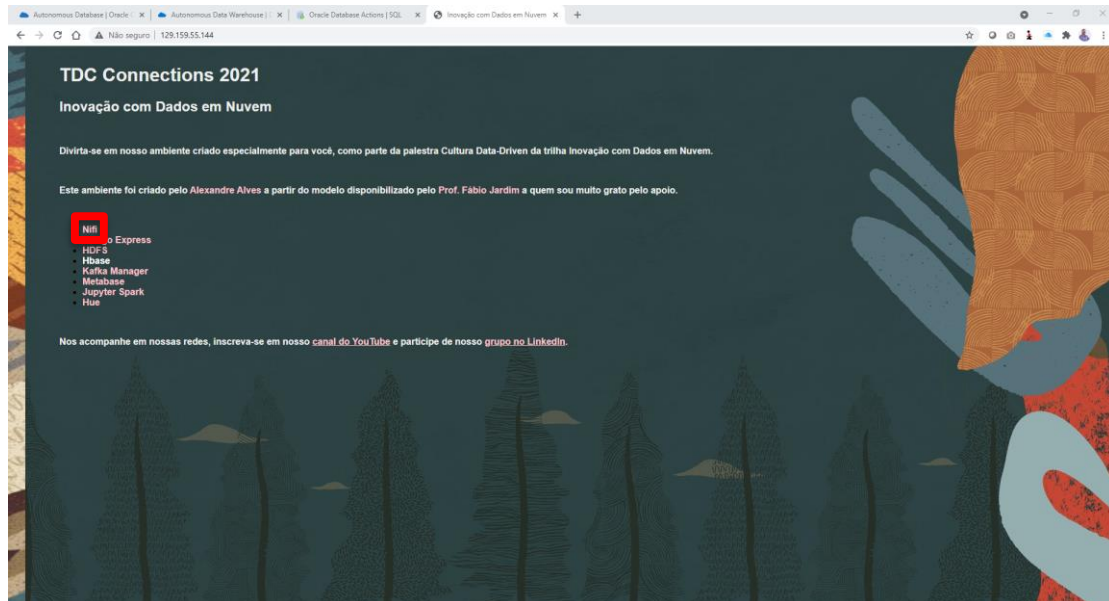
Como você pode ver acima, este ambiente tem diversos recursos para Big Data instalados e para este laboratório utilizaremos o NiFi, o MySQL e o Metabase.

Lab. 4 – Criando a ingestão com NiFi

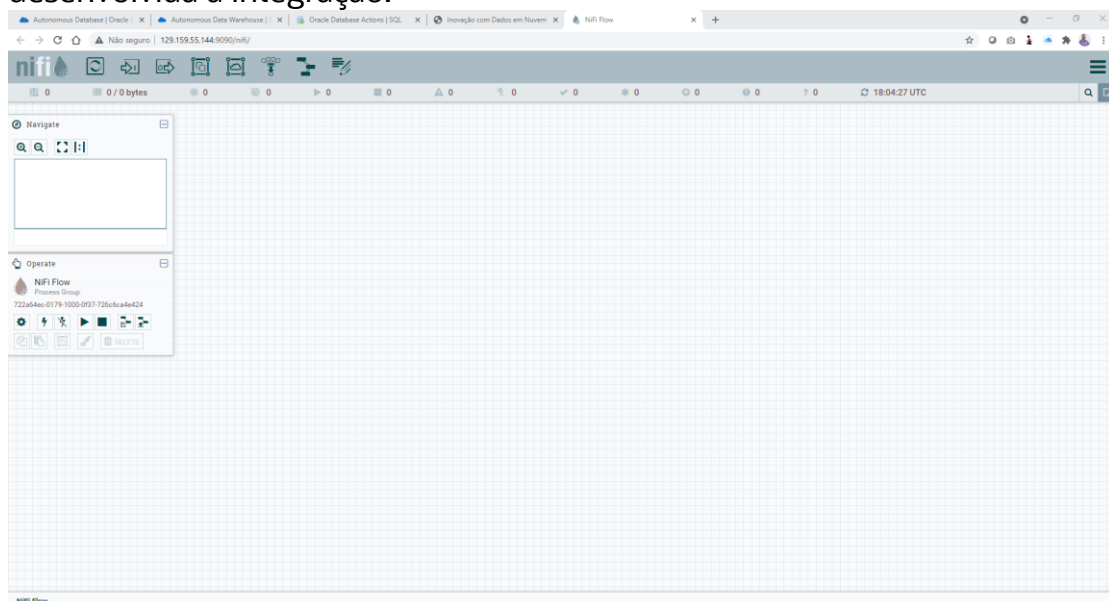
Objetivos:

- Utilizar o NiFi para ingerir dados do Twitter em tempo real para o Autonomous Database

Volte para a página de seu servidor web e clique no link para o **Nifi**:



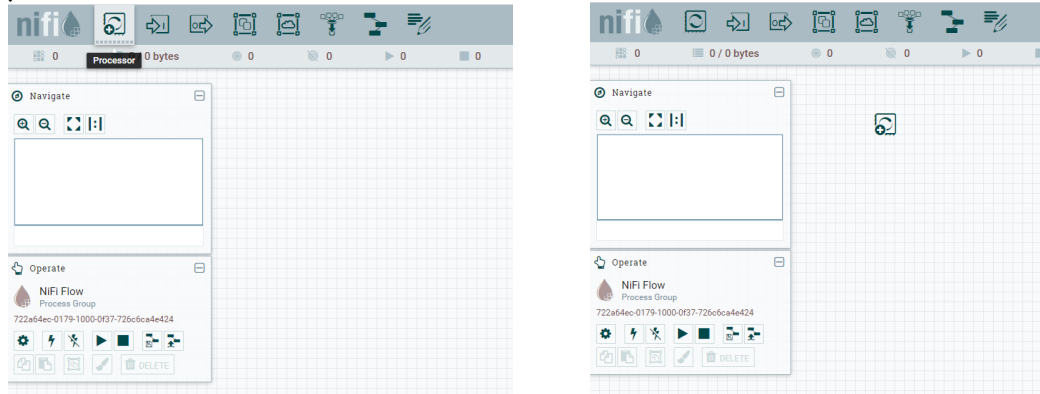
A interface do Nifi será aberta como na janela abaixo. A partir dela será desenvolvida a integração.



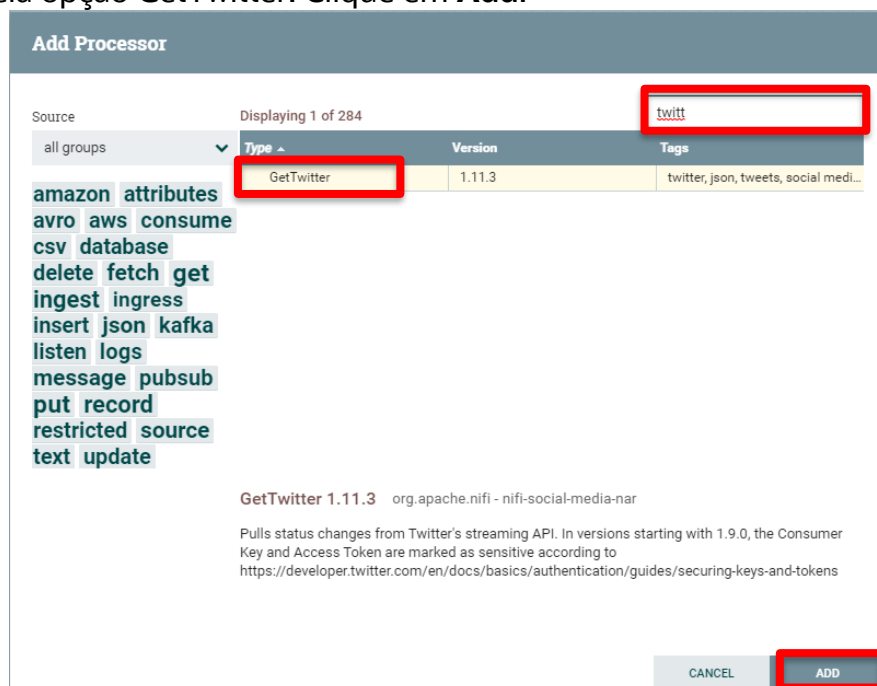
O NiFi é uma ferramenta open source criada pela NSA para fluxos de integração em tempo real. Cada processador é responsável por uma parte da

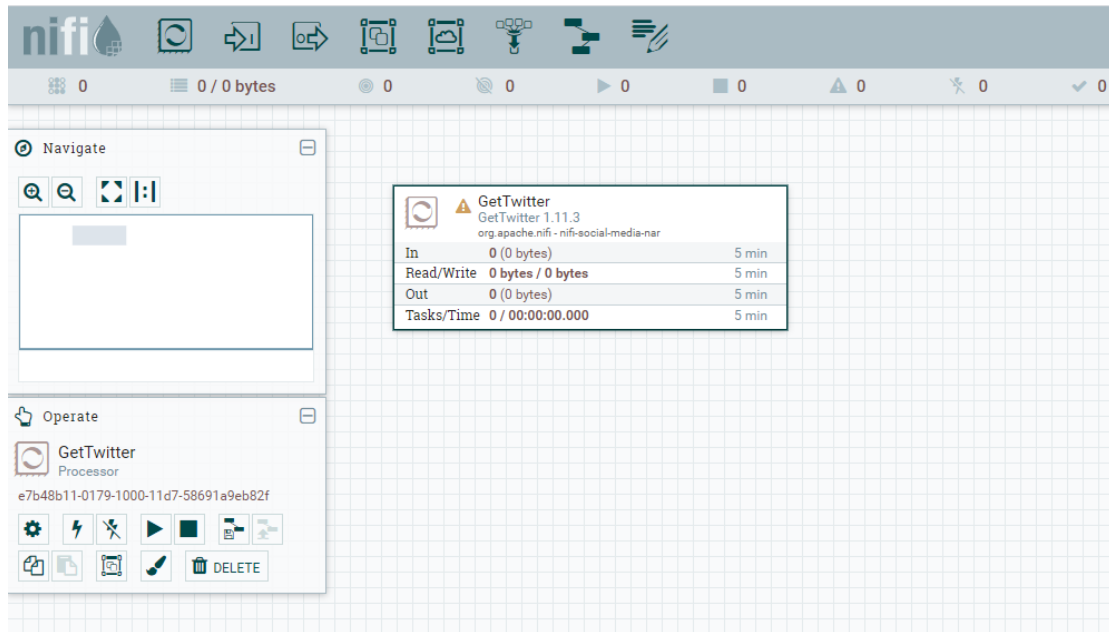
integração e baseado em uma arquitetura desacoplada, onde cada componente pode trabalhar independente do próximo, enfileirando as mensagens caso o próximo passo esteja indisponível.

Para iniciar o nosso fluxo clique no processor e arraste para uma área livre da janela.



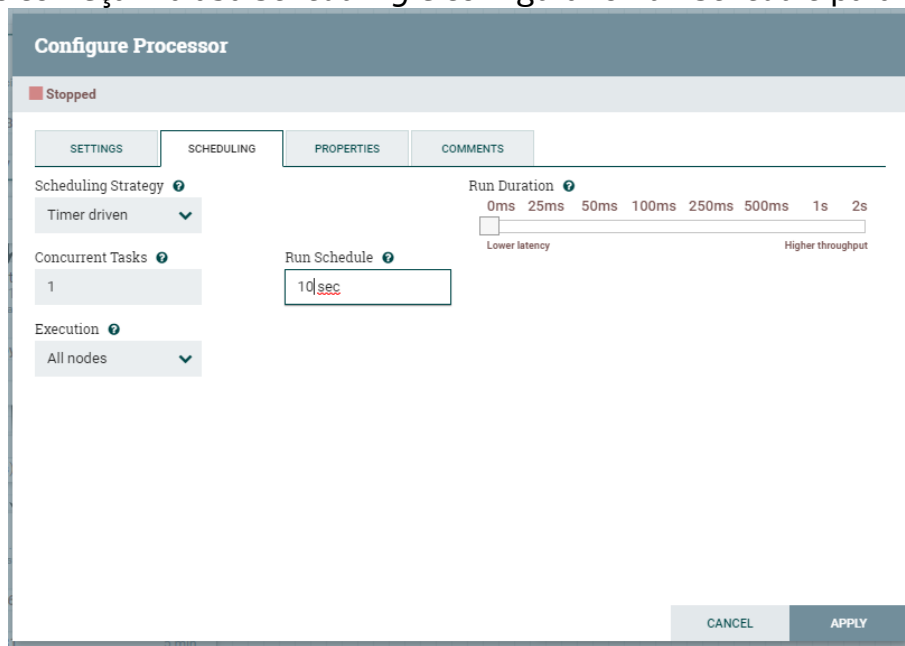
As inúmeras opções disponíveis para *processors* são exibidas. Procure no filtro pela opção GetTwitter. Clique em **Add**.





Dê um duplo-clique no novo *processor* adicionado para as configurações.

Vamos começar na aba *Scheduling* e configurar o *Run Schedule* para **10 sec**



Na sequência vamos configurar a aba *Properties*. Os campos devem ser preenchidos com:

- **Twitter Endpoint:** Filter Endpoint
- **Consumer Key:** preencher com a API Key gerada no momento da criação da sua conta de desenvolvedor no Twitter
- **Consumer Secret:** usar o API Secret gerado na criação da conta
- **Access Token:** fornecer o Access Token criado anteriormente durante este laboratório
- **Access Token Secret:** usar o Access Token Secret

- **Languages:** pt
- **Terms to Filter On:** lista de palavras chave que você utilizará para pesquisa no Twitter. Para nosso exemplo buscaremos por *TDC Connections* mas você pode usar outras palavras. Caso utilize mais de um termo eles devem ser separados por vírgula.

Configure Processor

Invalid

SETTINGS
SCHEDULING
PROPERTIES
COMMENTS

Required field

Property	Value
Twitter Endpoint	Filter Endpoint
Max Client Error Retries	5
Consumer Key	Sensitive value set
Consumer Secret	Sensitive value set
Access Token	Sensitive value set
Access Token Secret	Sensitive value set
Languages	pt
Terms to Filter On	TDC Connections
IDs to Follow	No value set
Locations to Filter On	No value set

CANCEL
APPLY

Aplique as alterações efetuadas (botão **Apply**) e vamos adicionar um novo *Processor*.

Coloque-o abaixo do anterior e procure por **EvaluateJsonPath**. Clique em **Add**.

Add Processor

Source

all groups

amazon attributes
avro aws consume
csv database
delete fetch get
ingest ingress
insert json kafka
listen logs
message pubsub
put record
restricted source
text update

Displaying 25 of 284

json

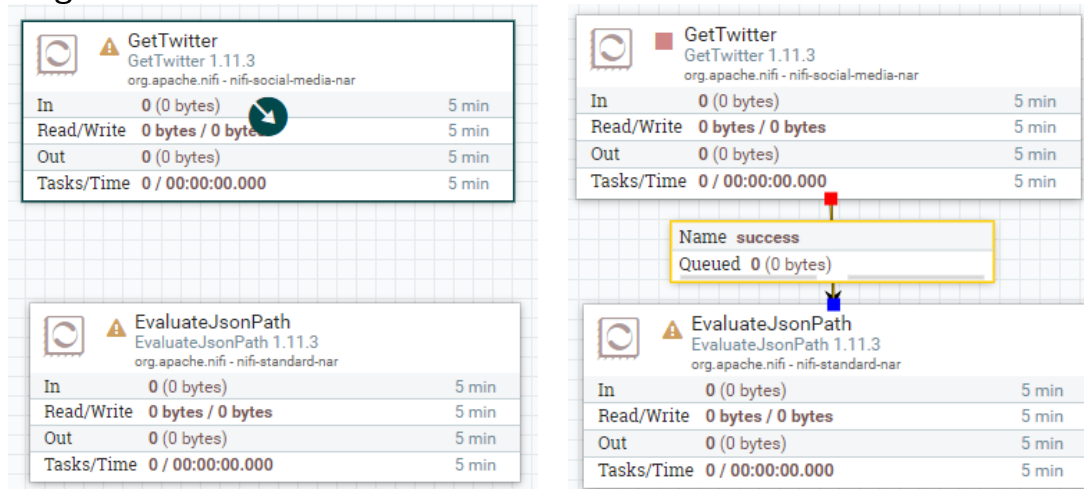
Type	Version	Tags
AttributesToJSON	1.11.3	flowfile, json, attributes
ConsumeKafkaRecord_0_10	1.11.3	0.10.x, PubSub, Consume, Inge...
ConsumeKafkaRecord_0_11	1.11.3	PubSub, Consume, Ingest, Get...
ConsumeKafkaRecord_1_0	1.11.3	PubSub, Consume, 1.0, Ingest, ...
ConsumeKafkaRecord_2_0	1.11.3	PubSub, Consume, Ingest, 2.0, ...
ConvertAvroToJSON	1.11.3	json, convert, avro
ConvertJSONToSQL	1.11.3	database, rdbms, flat, json, ins...
ConvertRecord	1.11.3	schema, log, record, csv, freefo...
EvaluateJsonPath	1.11.3	JSON, JsonPath, evaluate
FlattenJson	1.11.3	flatten, json
GetMongoRecord	1.11.3	mongo, get, fetch, record, json, ...
GetTwitter	1.11.3	twitter, json, tweets, serial ma...

EvaluateJsonPath 1.11.3 org.apache.nifi - nifi-standard-nar

Evaluates one or more JsonPath expressions against the content of a FlowFile. The results of those expressions are assigned to FlowFile Attributes or are written to the content of the FlowFile itself, depending on configuration of the Processor. JsonPaths are entered by adding user-defined properties; the name of the property maps to the Attribute Name into which the result will be plac...

CANCEL
ADD

Vamos ligar os fluxos. Coloque o mouse em cima do **GetTwitter** e ao aparecer a seta clique e arraste até o **EvaluateJsonPath**. Clique **Add** na janela que surge



Abra o **EvaluateJsonPath** para fazermos as configurações.

Na aba **Settings** vamos definir que o fluxo só deve seguir se as condições forem as esperadas (ou seja, retornar o campo *text* no fluxo anterior). Para isso marque **failure** e **unmatched**.

The screenshot shows the 'Configure Processor' dialog for the 'EvaluateJsonPath' processor. The 'Settings' tab is active, displaying various configuration options. The 'Name' field is set to 'EvaluateJsonPath', and the 'Type' is 'EvaluateJsonPath 1.11.3'. The 'Bundle' is 'org.apache.nifi - nifi-standard-nar'. The 'Penalty Duration' is '30 sec' and the 'Yield Duration' is '1 sec'. The 'Bulletin Level' is set to 'WARN'. On the right, under 'Automatically Terminate Relationships', the 'failure' and 'unmatched' checkboxes are selected, while 'matched' is not. The 'failure' checkbox is highlighted with a red box, and the 'unmatched' checkbox is also highlighted with a red box. The 'Apply' button is visible at the bottom right.

Na aba **Properties** vamos adicionar um atributo para capturarmos o texto proveniente do fluxo anterior. Clique no botão + e adicione a propriedade **text** com valor **\$.text**. Clique em **Apply**.

Configure Processor

Invalid

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

Property	Value
Destination	flowfile-content
Return Type	auto-detect
Path Not Found Behavior	ignore
Null Value Representation	empty string
text	\$.text

CANCEL APPLY

Na sequência vamos preparar o comando *insert* para inserir as mensagens no banco de dados. Adicione um *processor* do tipo **ReplaceText**. O **EvaluateJsonPath** deve ser conectado a este novo objeto pelo relacionamento **matched**.

Na aba **Settings** marque a opção **failure** e na aba **Properties** adicione `insert into tweets values('$.1')` à propriedade **Replacement Value**.

Configure Processor

Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

Property	Value
Search Value	(?s)(^.*\$)
Replacement Value	insert into tweets values('\$.1')
Character Set	UTF-8
Maximum Buffer Size	1 MB
Replacement Strategy	Regex Replace
Evaluation Mode	Entire text
Line-by-Line Evaluation Mode	All

CANCEL APPLY

O último passo é gravar as linhas extraídas no MySQL que também faz parte deste ambiente. Adicione um *processor* do tipo **PutSQL** e conecte-o como sucessor do **ReplaceText** marcando a opção **success**.

Na aba **Settings** marque todas as opções do **Automatically Terminate Relationships**, visto que essa é o último componente de nossa integração.

Configure Processor

Stopped

SETTINGS SCHEDULING **PROPERTIES** COMMENTS

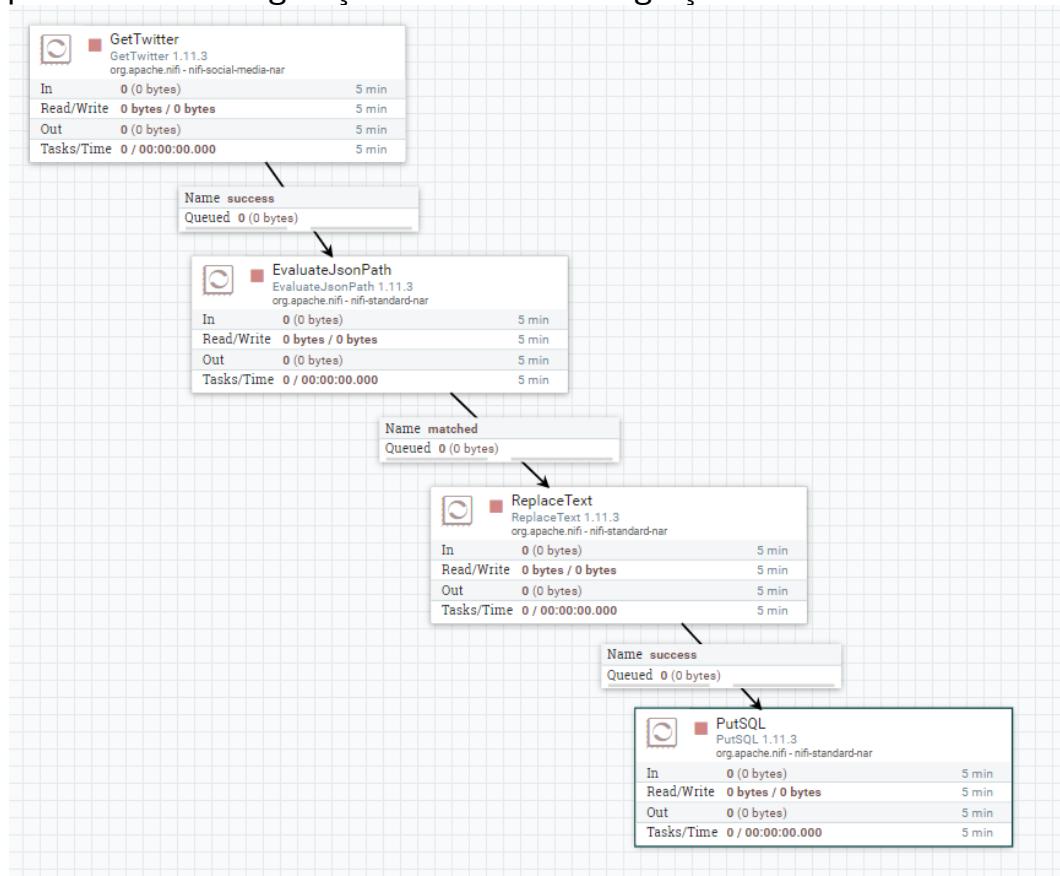
Required field +

Property	Value
JDBC Connection Pool	MySQL
SQL Statement	No value set
Support Fragmented Transactions	true
Database Session AutoCommit	false
Transaction Timeout	No value set
Batch Size	100
Obtain Generated Keys	false
Rollback On Failure	false

CANCEL APPLY

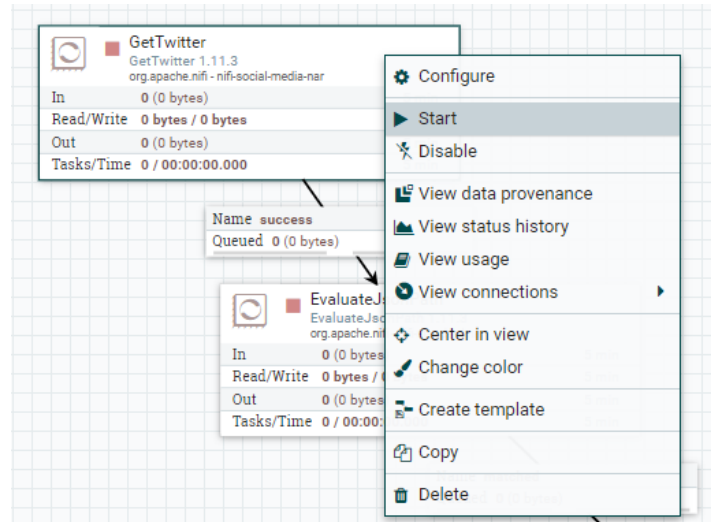
Na aba **Properties** só será necessário configurar **JDBC Connection Pool** com o valor **MySQL**, que já estava previamente criada para este ambiente.

Após todas as configurações feitas a sua integração deve ficar assim:

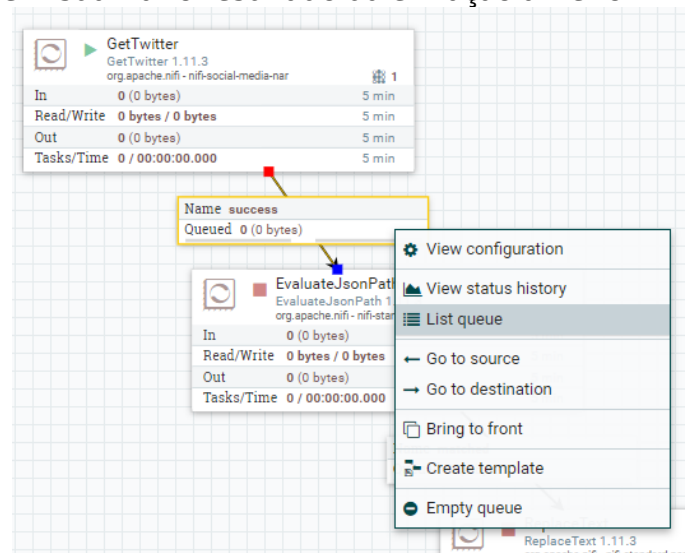


Conforme explicado anteriormente a arquitetura do NiFi é desacoplada, ou seja, cada *processor* tem a capacidade de atuar separadamente dos demais, podendo assim ler as mensagens da fila a partir de sua conexão predecessora e gravando na fila de sua conexão sucessora.

Para que os *processors* iniciem suas tarefas é necessário clicar com o botão direito do mouse em cada um deles e clicar em **Start**.



Inicie um processor por vez e observe como a fila do sucessor é preenchida. Você pode visualizar as mensagens que estão chegando à fila clicando com o botão direito do mouse nela e em **List queue**. Uma nova janela será aberta onde é possível visualizar o resultado da extração anterior.



Clique no botão do “olho” na mensagem para visualizar seu conteúdo.

SUCCESS

Displaying 4 of 4 (32.39 KB)

The source of this queue is currently running. This listing may no longer be accurate.

Position	UUID	Filename	File Size	Queue Duration	Lineage Duration	Penalized	
1	0788ebaf-5622-4093-804f-420601912331	0788ebaf-5622-4093-804f-420601912331.json	7.52 KB	00:00:43.479	00:00:43.479	No	✓
2	84dec909-6279-4828-9759-6a45df51ba54	84dec909-6279-4828-9759-6a45df51ba54.json	11.79 KB	00:00:33.469	00:00:33.469	No	✓
3	96354d73-1492-476d-b18a-0ba6b81a6efc	96354d73-1492-476d-b18a-0ba6b81a6efc.json	5.89 KB	00:00:23.468	00:00:23.468	No	✓
4	6a39ba02-959e-4123-b16e-3279c73ca5dc	6a39ba02-959e-4123-b16e-3279c73ca5dc.json	7.19 KB	00:00:13.467	00:00:13.467	No	✓

Last updated: 04:22:50 UTC

Uma vez que todos os *processors* estejam funcionando a ingestão seguirá acontecendo em segundo plano, mesmo que você feche a tela do NiFi.

Você pode observar os registros inseridos através do **Metabase**. O link para a ferramenta está disponível na página inicial de seu servidor.

TDC Connections 2021

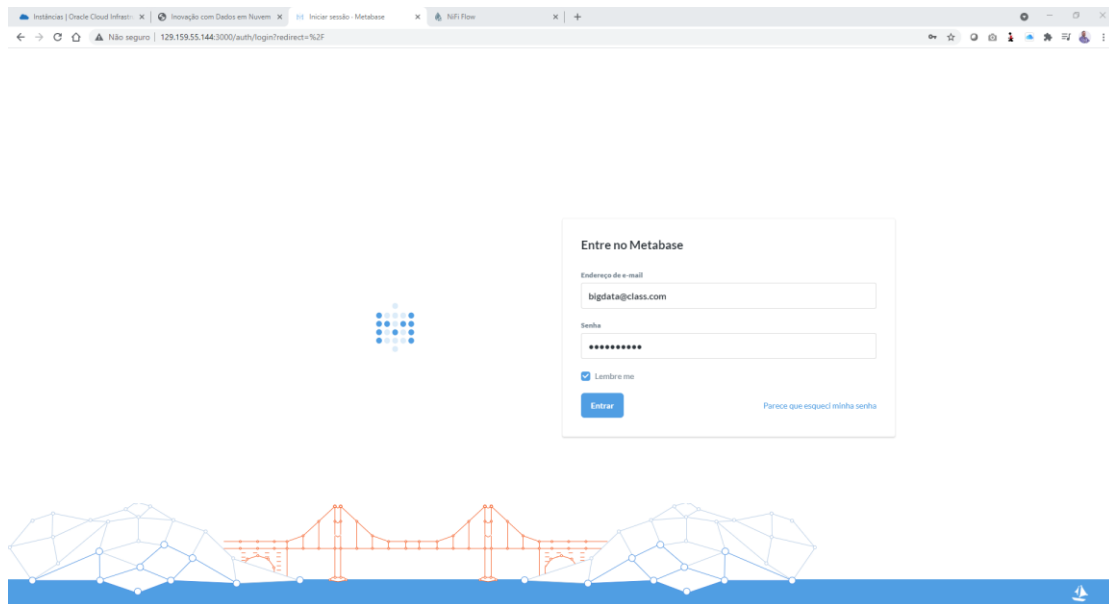
Inovação com Dados em Nuvem

Divirta-se em nosso ambiente criado especialmente para você, como parte da palestra Cultura Data-Driven da trilha Inovação com Dados em Nuvem.

Este ambiente foi criado pelo **Alexandre Alves** a partir do modelo disponibilizado pelo Prof. **Fábio Jardim** a quem sou muito grato pelo apoio.

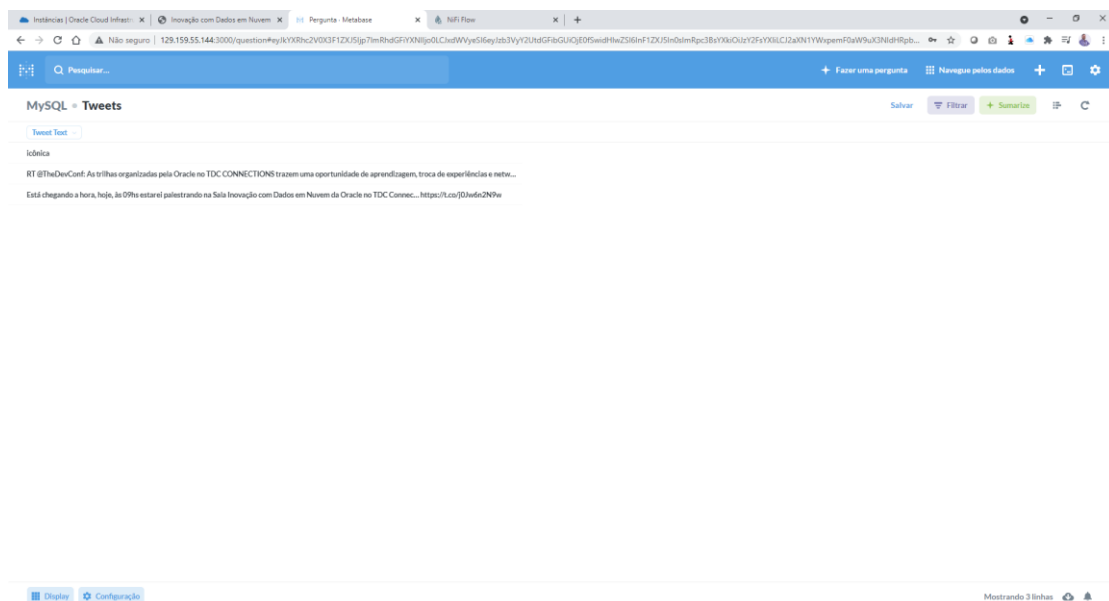
- NiFi
- Mongo Express
- HDFS
- Hbase
- **Kafka Manager**
- **Metabase**
- Jupyter Spark
- Hue

Nos acompanhe em nossas redes, inscreva-se em nosso [canal do YouTube](#) e participe de nosso [grupo no LinkedIn](#).



- Usuário: bigdata@class.com
- Senha: bigdata123

Na página inicial escolha o banco **MySQL** e na sequência tabela **Tweets**.



A partir deste ponto as mensagens dos usuários do Twitter passam a integrar sua solução de Big Data, podendo ser cruzada, adicionada, analisada e utilizada por suas equipes para saber o que as pessoas estão pensando de sua empresa, de sua nova estratégia de marketing ou como a concorrência tem atuado.

Parabéns!

Você completou o Laboratório
Hands-On de Ingestão de Dados

