

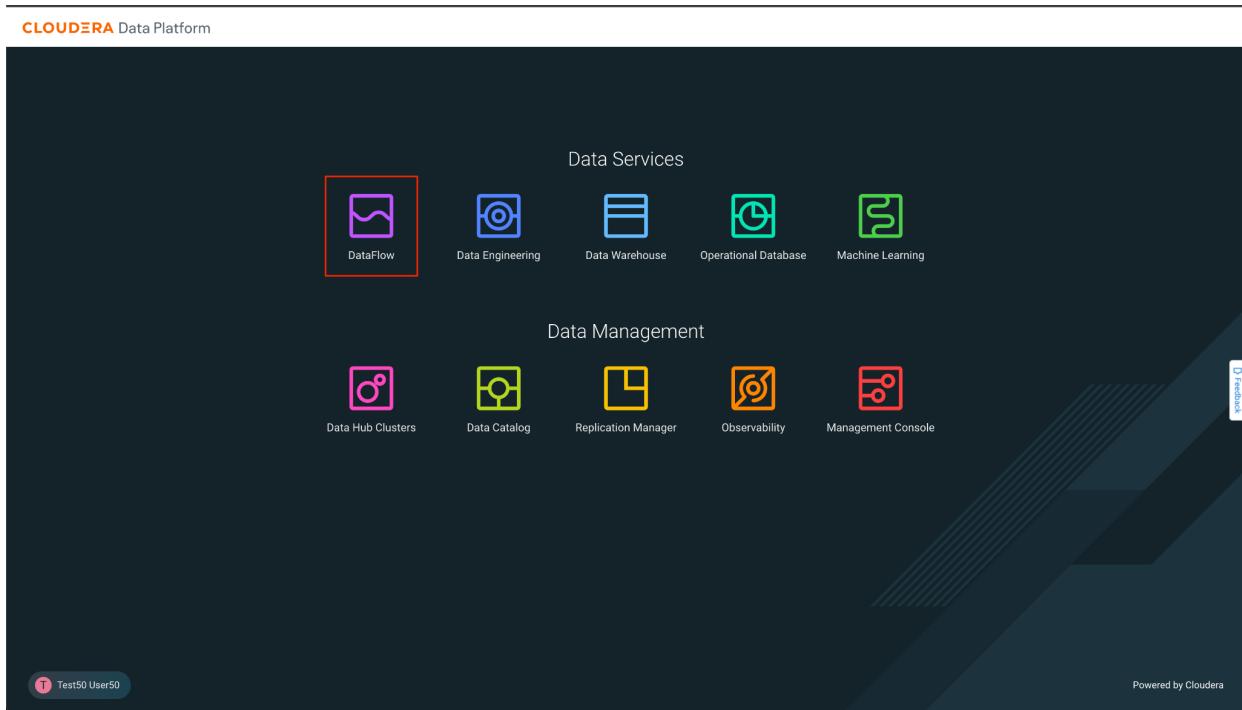
Ciclo de vida dos dados em CDP Public Cloud

Laboratório Data Flow

Metas:

- Consumir dados de um tópico Kafka
- Converta os dados para o formato Parquet
- Armazene os dados em uma tabela no Lakehouse

1. Clique em DataFlow no CDP PC Home:



2. Uma vez no DataFlow, clique na opção **Catalog** no menu à esquerda. Os modelos de aplicativo de ingestão de dados estão listados aqui. Para este workshop criamos e publicamos um modelo que permite ler os dados do tópico Kafka e ingeri-los/armazená-los no Lakehouse fornecido pelo CDP Public Cloud. Clique no fluxo chamado **kafka_to_lakehouse** para começar a implantá-lo.

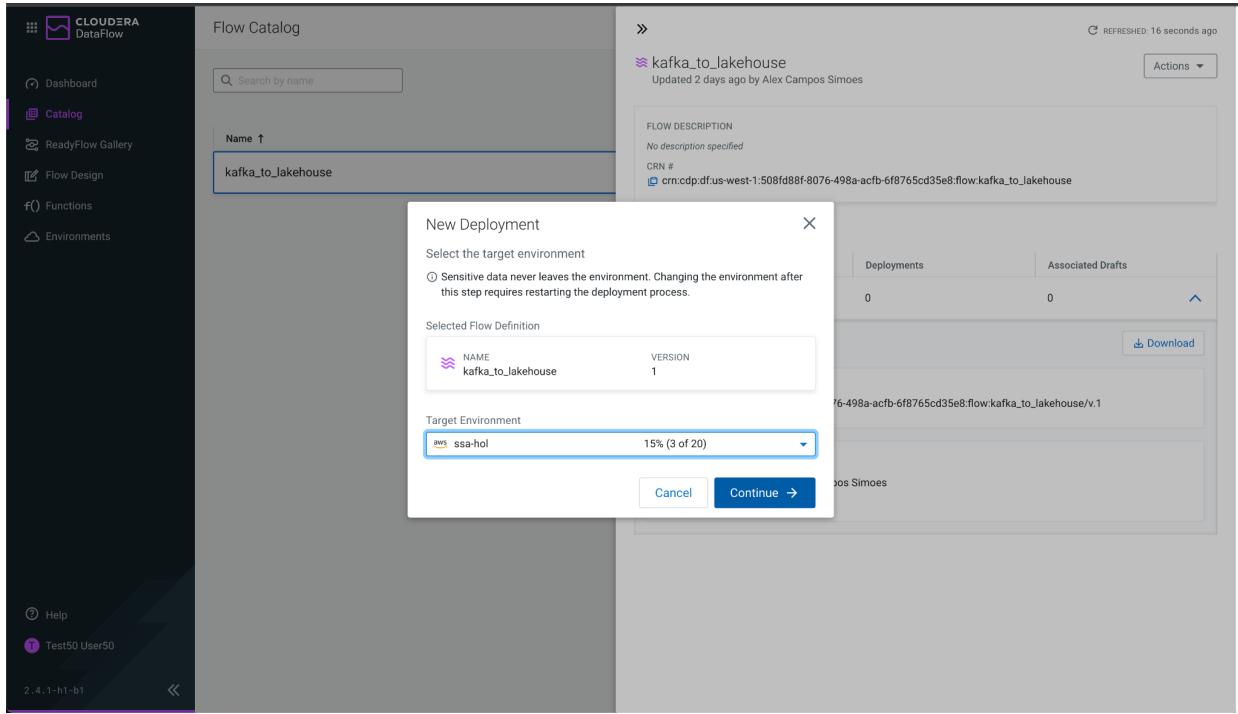
The screenshot shows the Cloudera DataFlow interface. On the left is a dark sidebar with navigation links: Dashboard, Catalog (which is selected and highlighted in purple), ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and Test50 User50. Below the sidebar, the version number is listed as 2.4.1-h1-b1. The main content area is titled "Flow Catalog". It features a search bar labeled "Search by name". A table lists a single flow entry: "kafka_to_lakehouse" (Type: Custom Flow Definition, Version: 1, Last Updated: 2 days ago). At the top right, there are buttons for "Import Flow Definition" and "REFRESHED: 5 seconds ago". At the bottom right, there are pagination controls: "Items per page: 10", "1 – 1 of 1", and navigation arrows.

3. Ao clicar o seguinte painel aparece com as informações do Fluxo. Mostra as versões disponíveis, data de criação, usuário de criação e um botão **Deploy** para iniciar a implantação. Clique nesse botão.

This screenshot shows the same Cloudera DataFlow interface, but the "kafka_to_lakehouse" flow entry from the previous screen is now selected and expanded. The expanded view includes:

- Flow Description:** No description specified.
- CRN #:** crn:cdp:df:us-west-1:508fd88f-8076-498a-acfb-6f8765cd35e8:flow:kafka_to_lakehouse
- Deployment Status:** Only show deployed versions (checkbox)
- Deployment Table:** Shows one version (1) with 0 deployments and 0 associated drafts.
- Deployment Buttons:** Deploy → (highlighted in blue) and Download.
- Deployment Details:** CRN # (same as above), CREATED (2023-05-19 00:15 CEST by Alex Campos Simoes), and a note: "Initial Version".

4. A janela pop-up a seguir permite selecionar o cluster DataFlow no qual você deseja implantar o fluxo. Neste caso o cluster é selecionado **ssa-hol**. O instrutor do workshop lhe dirá qual ambiente selecionar. Depois de selecionado, clique **Continue**.



5. A partir deste ponto, você precisará inserir as informações do fluxo. Comece atribuindo um nome (**Deployment Name**) e clique em **Next**.

*Para este workshop nomeie o fluxo com o nome de usuário atribuído, no caso do exemplo abaixo: **user050**.*

New Deployment

Overview

Deployment Name
user050
Deployment name is valid

Selected Flow Definition
NAME: kafka_to_lakehouse
VERSION: 1

Target Environment
aws ssa-hol

[Cancel](#) [Next →](#)

6. Confirme que a opção **Automatically start flow upon successful deployment** esteja marcada (com o check) e clique em **Next**.

New Deployment

NiFi Configuration

NiFi Runtime Version
CURRENT VERSION: Latest Version (1.23.2.2.3.10.0-23)
[Change Version](#)

Review the Cloudera DataFlow and CDP Runtime support matrix to ensure the selected NiFi Runtime Version is compatible.

Autostart Behavior
 Automatically start flow upon successful deployment

Inbound Connections
 Allow NiFi to receive data

Custom NAR Configuration
 This flow deployment uses custom NARs

Overview

FLOW DEFINITION: kafka_to_lakehouse v.1
ENVIRONMENT DEPLOYING TO: ps-sandbox-aws
DEPLOYMENT NAME: user050

[Cancel](#) [Previous](#) [Next →](#)

7. Nesta parte dos Parâmetros, você deve inserir os seguintes valores:

CDP Workload User Password: Digite a senha do usuário de serviço configurada para o usuário atribuído no início do workshop.

CDP Workload Username: insira o número de usuário atribuído, *user050* no exemplo abaixo.

Database: insira o número de usuário atribuído, *user050* no exemplo abaixo. Este banco de dados e as tabelas já estão pré-criados. Vamos analisá-lo mais tarde.

Kafka Consumer Group id: insira um valor exclusivo usando o usuário atribuído, *consumer_user050* no exemplo abaixo.

Verifique se os parâmetros foram inseridos corretamente. Em seguida, clique em **Next**.

New Deployment

Overview

NiFi Configuration

Parameters

Data entered here never leaves the environment in your cloud account.
Provide parameter values directly in the text input or upload a file for parameters that expect a file.

The selected flow definition references an external Default NiFi SSL Context Service. Hence, DataFlow will automatically create a matching SSL Context Service with a keystore and truststore generated from the target environment's FreeIPA certificate.

SHOW: Sensitive No value

parameters (7)

CDP Workload User Password
.....

CDP Workload Username
user050

CDPEnvironment

core-site.xml
ssl-client.xml
hive-site.xml

Select File
Drop file or browse

FLOW DEFINITION kafka_to_lakehouse v.1
ENVIRONMENT DEPLOYING TO ssa-hol
DEPLOYMENT NAME user050

NiFi Configuration

NIFI RUNTIME VERSION Latest Version (1.20.0.2.3.8.2-2)
AUTO-START FLOW No
INBOUND CONNECTIONS No
CUSTOM NAR CONFIGURATION No

Cancel ← Previous Next →

This screenshot shows the 'New Deployment' interface in a cloud-based environment. The left sidebar shows a navigation path: Overview (selected), NiFi Configuration (selected), Parameters (highlighted), Sizing & Scaling, Key Performance Indicators, and Review. The main area is titled 'Parameters' and contains a note about entering sensitive data. It shows two parameter entries: 'CDP Workload User Password' (with a redacted value) and 'CDP Workload Username' (set to 'user050'). Below these is a 'CDPEnvironment' section with three XML files listed: 'core-site.xml', 'ssl-client.xml', and 'hive-site.xml', each with a radio button next to it. A 'Select File' button is also present. To the right, a summary panel for 'NiFi Configuration' displays the flow definition 'kafka_to_lakehouse v.1', the deployment environment 'ssa-hol', and the deployment name 'user050'. At the bottom, there are 'Cancel', '← Previous', and 'Next →' buttons.

New Deployment

CDPEnvironment

- core-site.xml
- ssl-client.xml
- hive-site.xml

Select File
Drop file or browse

Database

Kafka Brokers

Kafka Consumer Group Id

Kafka Topic

NiFi Configuration

Flow Definition: kafka_to_lakehouse v.1

Environment Deploying To: ssa-hol

Deployment Name: user050

Auto-Start Flow: No

Inbound Connections: No

Custom NAR Configuration: No

Cancel ← Previous Next →

8. Neste workshop não vamos alterar os parâmetros de sizing e scaling automático, então clique em **Next**.

New Deployment

Sizing & Scaling

Select the NiFi node size and the number of nodes provisioned for your flow.

NiFi Node Sizing

- Extra Small
- Small
- Medium
- Large

Number of NiFi Nodes

Auto Scaling

Nodes: 1

NiFi Configuration

Flow Definition: kafka_to_lakehouse v.1

Environment Deploying To: ssa-hol

Deployment Name: user050

Auto-Start Flow: No

Inbound Connections: No

Custom NAR Configuration: No

Parameters

parameters

CDP WORKLOAD USER PASSWORD
[Sensitive Value Provided]

CDP WORKLOAD USERNAME
user050

CDPENVIRONMENT

core-site.xml

ssl-client.xml

hive-site.xml

DATABASE

user050

KAFKA BROKERS

realtime-ingestion-corebroker0.ssa-hol.yu1t-vbzg.cloudera.site:9093,realtime-ingestion-corebroker1.ssa-hol.yu1t-vbzg.cloudera.site:9093,realtime-ingestion-corebroker2.ssa-hol.yu1t-vbzg.cloudera.site:9093

Cancel ← Previous Next →

9. Também não vamos configurar KPIs, então clique em **Next**.

New Deployment

Key Performance Indicators

Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed.

[Learn more](#)

[Add New KPI](#)

Cancel Previous Next →

10. Revise todas as informações fornecidas e então clique em **Deploy** para iniciar o processo de implantação.

New Deployment

Review

Overview

FLOW DEFINITION
kafka_to_lakehouse_v1

ENVIRONMENT DEPLOYING TO
ssa-hol

DEPLOYMENT NAME
user050

NiFi Configuration

NIFI RUNTIME VERSION
Latest Version (1.20.0.2.3.8.2-2)

AUTO-START FLOW
No

INBOUND CONNECTIONS
No

CUSTOM NAR CONFIGURATION
No

Parameters

parameters

CDP WORKLOAD USER PASSWORD
[Sensitive Value Provided]

CDP WORKLOAD USERNAME
user050

CDPENVIRONMENT

core-site.xml
ssl-client.xml
hive-site.xml

DATABASE

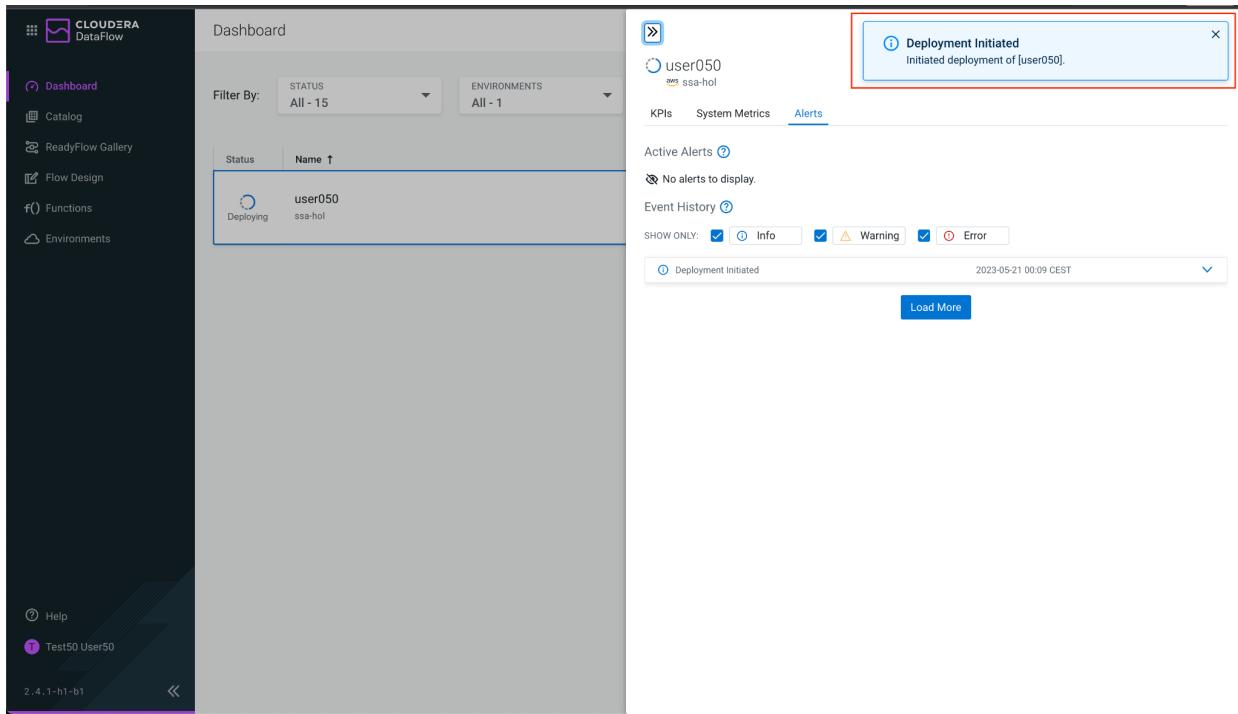
user050

KAFKA BROKERS

realtime-ingestion-corebroker0.ssa-hol.yu1t-vbzg.cloudera.site:9093,realtime-ingestion-corebroker1.ssa-hol.yu1t-vbzg.cloudera.site:9093,realtime-ingestion-corebroker2.ssa-hol.yu1t-vbzg.cloudera.site:9093

Cancel Previous Deploy

11. A caixa de pop-up azul indica que o processo de implantação do Flow foi iniciado. Ao clicar no botão **Load More** você poderá ver os diferentes estágios da implantação. Após cerca de 60 a 90 segundos, o último evento deve ser **Deployment Successful**.



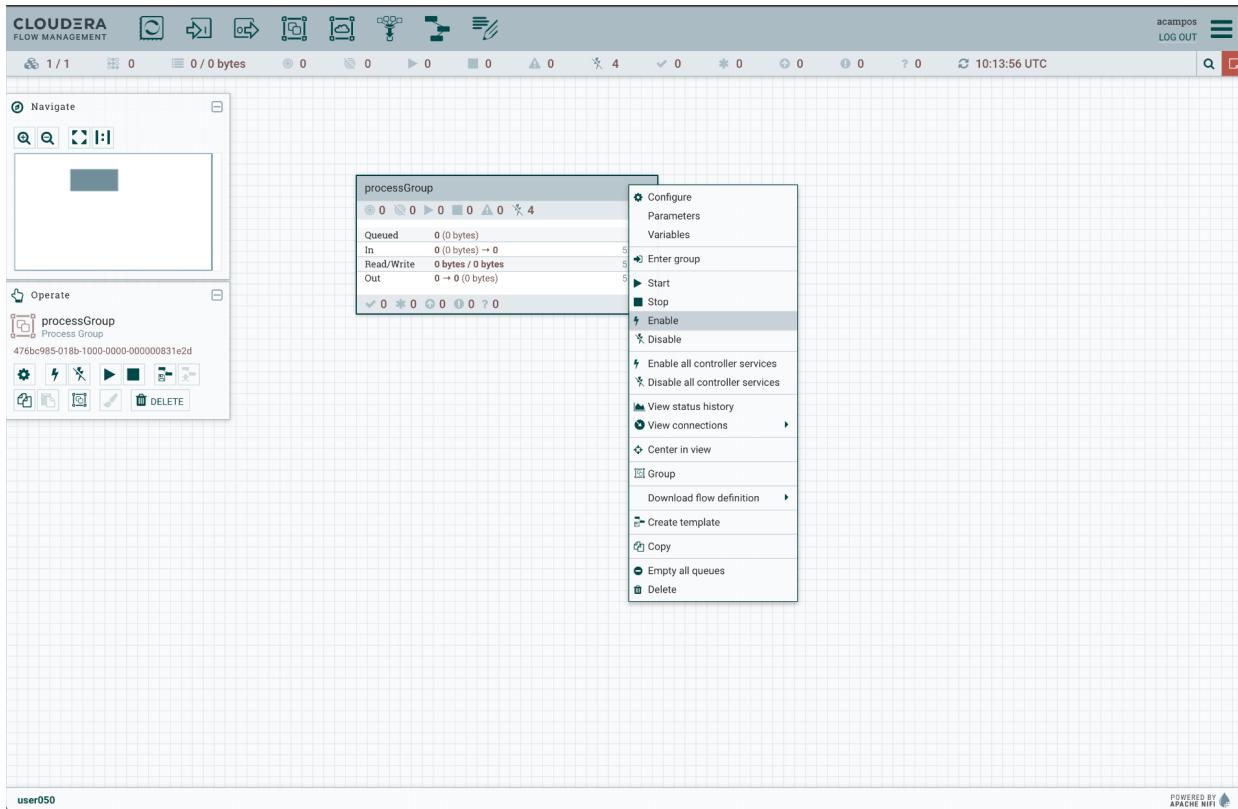
12. Após a conclusão da implantação, clique em **Manage Deployment** para ver os detalhes do Flow implantado.

The screenshot shows the Cloudera DataFlow interface. On the left, there's a sidebar with navigation links: Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and Test50 User50. The main area has a header "Dashboard" and a sub-header "user050 ssa-hol". It includes filters for "Filter By: STATUS All - 15" and "ENVIRONMENTS All - 1". Below this is a table with columns "Status" and "Name ↑", showing one entry: "user050 Deploying ssa-hol". To the right, there's a "KPIs" tab, a "System Metrics" tab, and an active "Alerts" tab. The "Alerts" tab shows a message "No alerts to display." and an "Event History" section with a table of log entries. A red box highlights the "Manage Deployment" button at the top right of the alerts section.

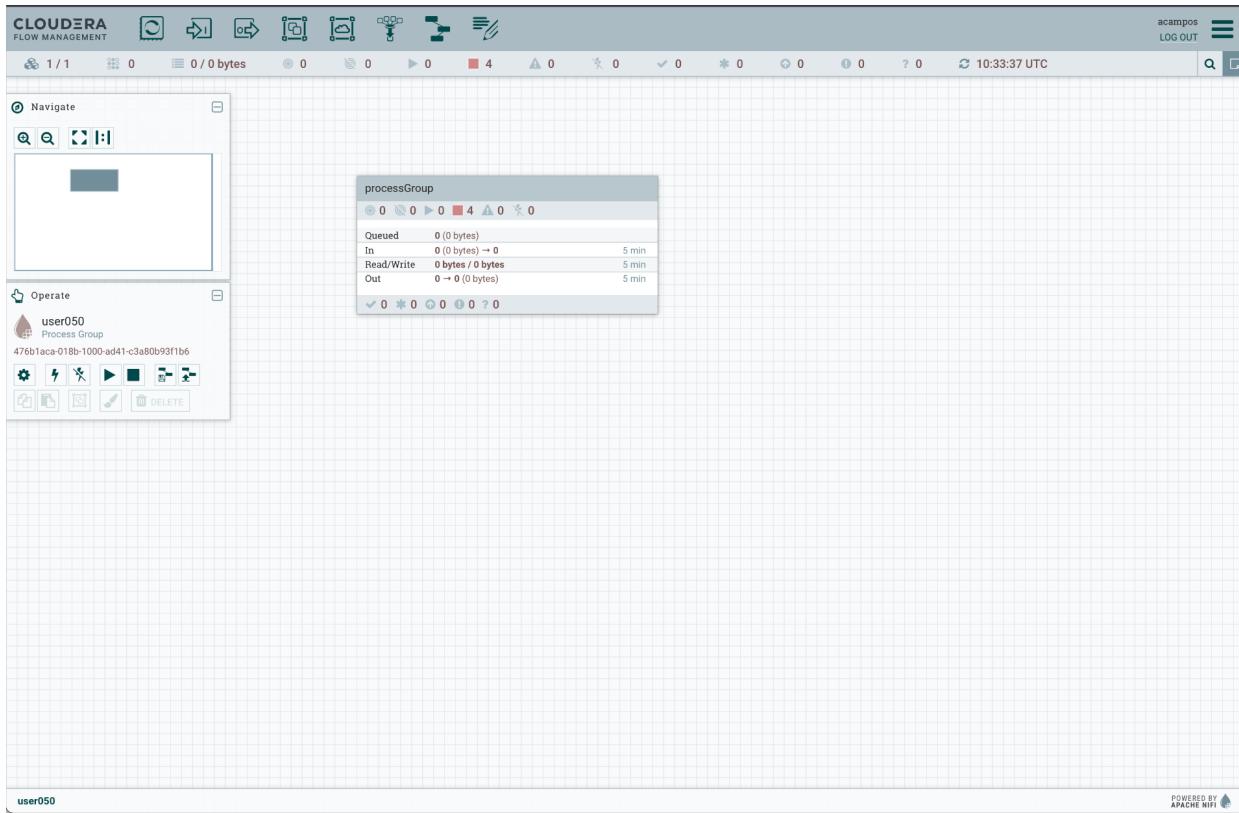
13. Nesta janela, você verá as informações gerais de Flow implantado. É hora de executar os processos do a partir da interface gráfica. Clique em **Actions** ->**View in NiFi**. Em outra janela será aberta a tela do Cloudera Flow Management.

The screenshot shows the "Deployment Manager" page for deployment "user050". It displays deployment details: Status (Deployed), Deployment Name (user050), Flow Definition (kafka_to_lakehouse V.1), Node Count (1), Auto Scaling (Disabled), Environment (aws ssa-hol), Region (US East(N. Virginia)), and NIIFI Runtime Version (1.20.0.2.3.8.2). On the right, there's an "Actions" dropdown menu with options: View in NiFi, Start flow, Change NiFi Runtime Version, Restart Deployment, and Terminate. Below the details, there's a "Deployment Settings" section with tabs for KPIs and Alerts, Sizing and Scaling, Parameters, and NIIFI Configuration. A "Key Performance Indicators" section allows users to add new KPIs. At the bottom, there are "Discard Changes" and "Apply Changes" buttons, and a link to "Update Deployment CLI Command".

14. Na nova janela, você poderá ver a tela Flow Management com um grupo de processos (uma caixa) chamado **processGroup**. Primeiro você precisa habilitar o grupo de processos. Clique com o botão direito no Process Group e selecione a opção **Enable**.



15. Em seguida, clique duas vezes no **processGroup** para abrir-lo.



16. Ao abrir o process group, são exibidos os processors que compõem o fluxo. Para resumir, existem quatro processors:

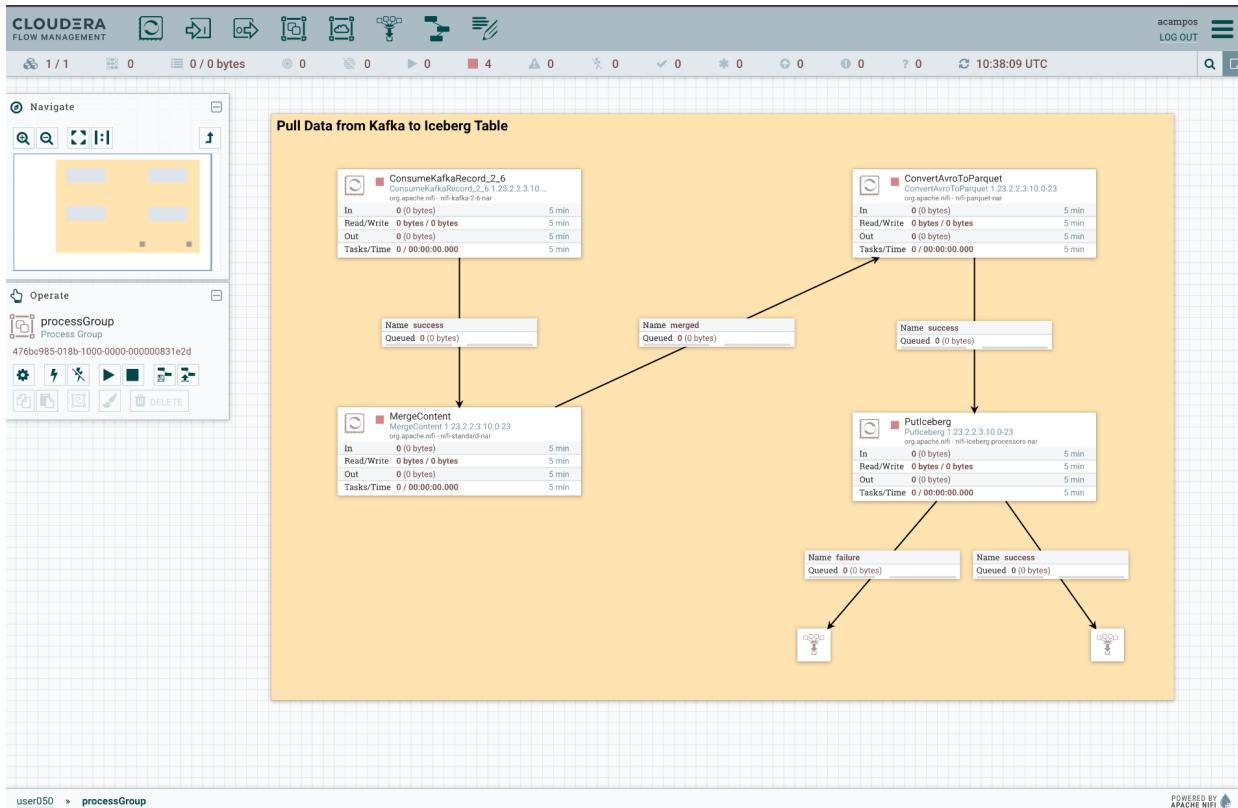
ConsumeKafkaRecord, processador para consumir dados do tópico Kafka, lendo os dados no formato JSON e gerando a saída no formato AVRO.

MergeContent, para agrupar os arquivos simplificando o fluxo de dados.

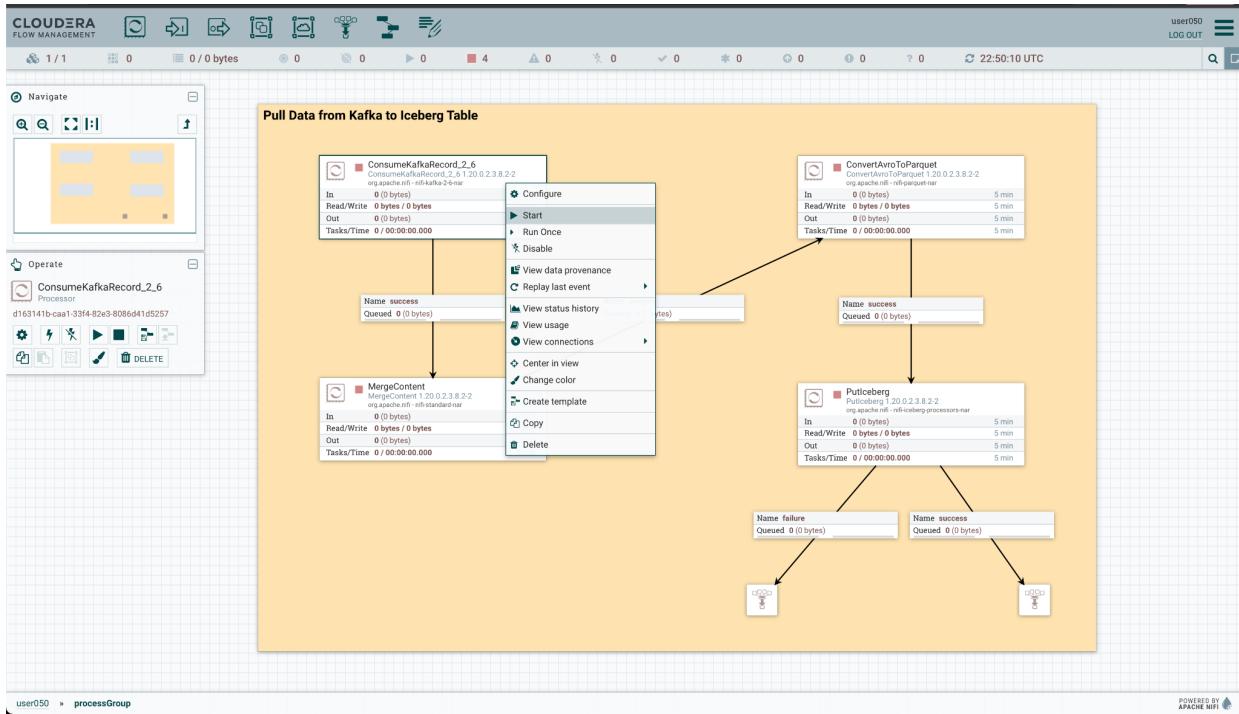
ConvertAvroToParquet, conversão necessária para salvar os dados no formato PARQUET.

PutIceberg, para inserir os dados na tabela no Lakehouse. A tabela de destino é chamada **telco_kafka_iceberg** e cada usuário tem um banco de dados atribuído.

Como pode ver, os processadores não são iniciados, estão pausados.

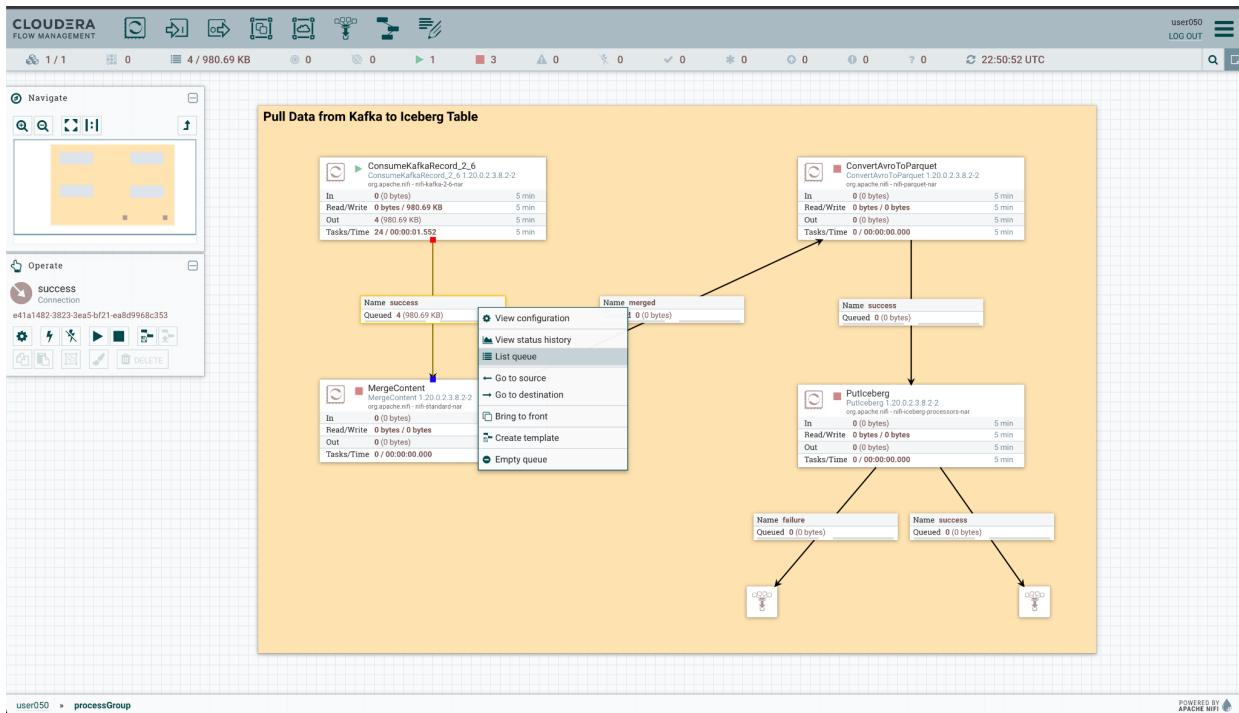


17. Agora já podes iniciar o pipeline. Comece com **ConsumeKafkaRecord**, clicando com o botão direito nele e depois clicando em **Start**. Isso começará a consumir os dados do tópico Kafka.



18. O Flow Management nos permite ver e acessar dados em movimento durante a execução do fluxo de dados. Entre o processor **ConsumeKafkaRecord** (que acabamos de iniciar) e **MergeContent** há uma conexão. Essa conexão é o que une os processors transmitindo dados entre eles.

Para verificar quantos dados estão enfileirados nessa conexão, atualize o contador pressionando a combinação Ctrl+R (Windows) ou Command+R (Mac) no teclado. Isso permitirá que as métricas atuais de todo o fluxo de dados sejam atualizadas. Em algum momento deve haver um número ao lado da legenda **Queued** na ligação entre **ConsumeKafkaRecord** e **MergeContent**. Para ver os dados enfileirados, clique com o botão direito na conexão e clique na opção **Listar Queue**, abrindo uma janela pop-up.



19. A próxima janela pop-up lista os dados enfileirados. Clique no ícone de informação (i) que aparece no lado esquerdo para visualizar os eventos.

SUCCESS							
Displaying 4 of 4 (980.69 KB)							
Position	UU	Filename	File Size	Queued Duration	Lineage Duration	Penalized	Node
1	2055d337-695f-4c66-8203-3ece27a62d...	2055d337-695f-4c66-8203-3ece27a62d...	278.24 KB	00:00:12.787	00:00:13.068	No	dfx-nifi-0.dfx-nifi.dfx-user050.ns.svc.clu...
2	510c8074-9798-4199-a228-d7894aca9...	510c8074-9798-4199-a228-d7894aca9...	283.60 KB	00:00:11.664	00:00:11.733	No	dfx-nifi-0.dfx-nifi.dfx-user050.ns.svc.clu...
3	cad12e7c-e301-439c-85b3-a53fb0f13a2a	cad12e7c-e301-439c-85b3-a53fb0f13a2a	285.48 KB	00:00:11.575	00:00:11.647	No	dfx-nifi-0.dfx-nifi.dfx-user050.ns.svc.clu...
4	01ee7d33-8e54-4a2b-a39c-a3f965b3cf87	01ee7d33-8e54-4a2b-a39c-a3f965b3cf87	133.37 KB	00:00:11.527	00:00:11.567	No	dfx-nifi-0.dfx-nifi.dfx-user050.ns.svc.clu...

20. Quando a janela de detalhes do FlowFile aparecer, clique no botão **View** para abrir o conteúdo dos eventos consumidos.

CloudBees Data Center

success

Displaying 4 of 4 (980.69 KB)

Position	UUID	Filename	File Size	Queued Duration	Lineage Duration	Penalized	Node
1	2055d337-695f-4c6d-8203-3ece27a62de...	2055d337-695f-4c6d-8203-3ece27a62de...	278.24 KB	00:00:12.787	00:00:13.068	No	dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.clu...
2	510c8074-9798-4199-a228-ad7894ac0a9...	510c8074-9798-4199-a228-ad7894ac0a9...	283.60 KB	00:00:11.664	00:00:11.733	No	dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.clu...
3	cad12e7e-e301-439c-85b3-a53fb0f13a2a	cad12e7e-e301-439c-85b3-a53fb0f13a2a	285.48 KB	00:00:11.575	00:00:11.647	No	dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.clu...
4	01ee7d33-8e54-4a2b-a39c-a3f965b3cf87	01ee7d33-8e54-4a2b-a39c-a3f965b3cf87	01ee7d33-8e54-4a2b-a39c-a3f965b3cf87			No	dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.clu...

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

FlowFile

DETAILS

ATTRIBUTES

FlowFile Details

UUID	2055d337-695f-4c6d-8203-3ece27a62dee
Filename	2055d337-695f-4c6d-8203-3ece27a62dee
File Size	278.24 KB
Queue Position	No value set
Queued Duration	00:00:19.534
Lineage Duration	00:00:19.815
Penalized	No
Node Address	dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443

Content Claim

Container	default
Section	1
Identifier	1684623047700-1
Offset	0
Size	278.24 KB

Download View

OK

Last updated: 22:50:59 UTC

user050 > processGroup

21. A nova janela que se abre mostra os dados do conteúdo do FlowFile. Por estar no formato AVRO, não é totalmente legível. Um desserializador deve ser selecionado para exibir os dados corretamente. Para isso, no canto superior esquerdo, selecione a opção **formatted** do menu **View as**.

22. Agora você pode exibir os dados corretamente. Observe que aparecem os campos ou atributos indicados no início do workshop. Você pode fechar a janela do FlowFile e os pop-ups, retornando à tela com os quatro processadores.



The screenshot shows a terminal window with the following details:

- Icon: A blue water droplet icon.
- Title: View as: formatted
- File information: Filename: 2055d337-695f-4c6d-8203-3ece27a62dee, Content Type: application/avro-binary
- Content: The terminal displays the contents of an Avro file, which is a JSON-like structure of customer data. The data includes fields such as 'multipelines', 'paperlessbilling', 'gender', 'onlinesecurity', 'internetservice', 'techsupport', 'contract', 'tenure', 'seniorcitizen', 'deviceprotection', 'streamingmovies', 'partner', 'monthlycharges', 'dependents', 'onlinesbackup', 'phoneservice', 'totalcharges', 'monthlycharges', 'paymentmethod', and 'customerid'. The data is presented in a multi-line JSON format, with each line representing a different customer record.

23. Continue executando cada um dos processadores na seguinte ordem: **MergeContent**, **ConvertAvroToParquet** e finalmente **PutIceberg**. Lembre-se que você pode atualizar os contadores do fluxo com a combinação Control+R ou Command+R.

Se as etapas anteriores foram executadas corretamente, a conexão do processador **PutIceberg** a um funil deve ser do tipo **success**.

