

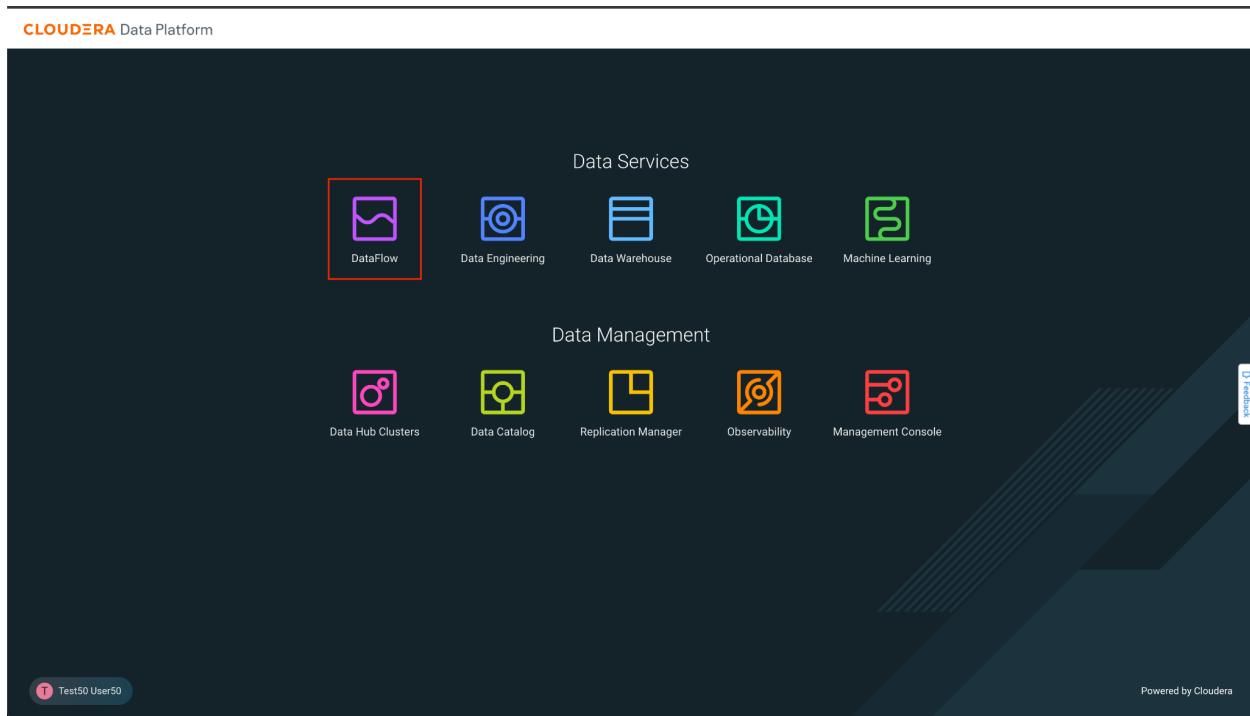
# Data Lifecycle CDP Public Cloud

## Data Flow Lab

Goals:

- Consume data from a Kafka topic
- Convert the data to Parquet format
- Store the data in a table in the Lakehouse

1. Click on DataFlow from CDP PC Home:



2. Once in DataFlow, click on the option **Catalog** from the left menu. The data ingestion application templates are listed here. For the purpose of this workshop, we have created and published a template that allows you to read Kafka topic data and ingest/store it in the Lakehouse provided by CDP Public Cloud. Click on the Flow called **kafka\_to\_lakehouse** to start deploying it.

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 a user profile for 'Test50 User50'. The main area is titled 'Flow Catalog' and contains a search bar labeled 'Search by name'. A table lists one flow entry:

Name	Type	Versions	Last Updated
kafka_to_lakehouse	Custom Flow Definition	1	2 days ago

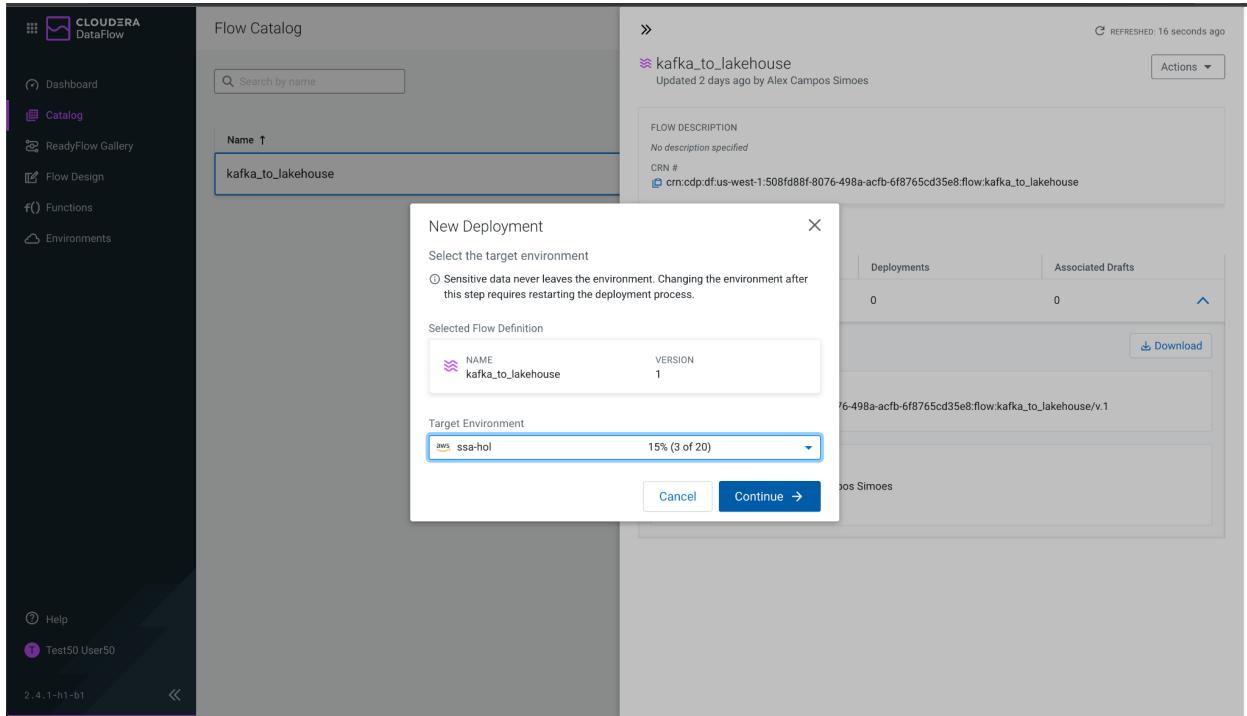
Below the table are buttons for 'Import Flow Definition' and 'REFRESHED: 5 seconds ago'. At the bottom, there are pagination controls: 'Items per page: 10', '1 – 1 of 1', and navigation arrows.

3. When clicked, the following panel appears with the Flow information. It shows the available versions, creation date, creator user, and a button **Deploy** to start the deployment. Click on that button.

This screenshot shows the same Cloudera DataFlow interface, but the 'kafka\_to\_lakehouse' flow is now selected in the catalog, indicated by a blue border around its row. The right-hand panel displays detailed information about this flow:

- 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)
- Associated Drafts:** 0
- Deployment Buttons:** Deploy → (highlighted in blue) and Download
- Created:** 2023-05-19 00:15 CEST by Alex Campos Simoes  
"Initial Version"

4. The following popup window allows you to select the DataFlow cluster in which you want to deploy the Flow. In this case, the cluster to be selected is **ssa-hol**. The workshop instructor will tell you which environment to select. Once selected, click **Continue**.



5. From this point, you will need to enter the Flow configuration. Start by assigning a name (**Deployment Name**) and click **Next**.

*For the purposes of this workshop, please name the Flow with the assigned username -user050, for example.*

New Deployment

Overview

Deployment Name  
user050  
Deployment name is valid

Selected Flow Definition

NAME	VERSION
kafka_to_lakehouse	1

Target Environment

NAME
aws
ssahol

[Cancel](#) [Next >](#)

6. Make sure the option **Automatically start flow upon successful deployment** is checked and click **Next**.

New Deployment

NiFi Configuration

NiFi Runtime Version [Change Version](#)

CURRENT VERSION  
Latest Version (1.23.2.2.10.0-23)

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

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

7. In this part of Parameters, you must enter the following values:

**CDP Workload User Password:** Enter the Workload Password shared at the beginning of the workshop.

**CDP Workload Username:** enter the assigned user number, *user050*, for example.

**Database:** enter the assigned user number, *user050*, for example. This database and the tables are already pre-created for you. We'll review it later.

**Kafka Consumer Group Id:** Enter a unique value using the assigned user. You can combine with the user id assigned for you.

Review that the parameters were entered correctly. Then click on **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

user050

CDP Workload Username

CDPEnvironment

core-site.xml

ssl-client.xml

hive-site.xml

0/100K

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 →

New Deployment

Overview

NiFi Configuration

Parameters

Sizing & Scaling

Key Performance Indicators

Review

**CDPEnvironment**

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

Drop file or browse

0/100K

DataFlow automatically adds all required configuration files to interact with Data Lake services. Unnecessary files that are added won't impact the deployment process.

**Database**

user050

7/100K

**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

203/100K

**Kafka Consumer Group Id**

Consumer\_user050

16/100K

**Kafka Topic**

telco\_data

10/100K

**Overview**

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 →**

8. There is no need to configure auto scaling parameters, then click on **Next**.

New Deployment

Overview

NiFi Configuration

Parameters

**Sizing & Scaling**

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

**NiFi Node Sizing**

Extra Small  
2 vCores Per Node  
4 GB Per Node

Small  
3 vCores Per Node  
6 GB Per Node

Medium  
6 vCores Per Node  
12 GB Per Node

Large  
12 vCores Per Node  
24 GB Per Node

**Number of NiFi Nodes**

Auto Scaling  Disabled

Nodes:

**Overview**

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

**Parameters**

parameters

COP WORKLOAD USER PASSWORD *[Sensitive Value Provided]*

COP WORKLOAD USERNAME user050

COPENVIRONMENT

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. We are also not going to configure KPIs by now, then click on **Next** to continue the configuration.

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](#)

**Overview**  
FLOW DEFINITION: kafka\_to\_lakehouse v.1  
ENVIRONMENT DEPLOYING TO: ssa-hol  
DEPLOYMENT NAME: user050

**NiFi Configuration**  
NIIFI 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](#) [Next](#)

10. Review all the information entered for your Flow, then click on **Deploy** to start the deployment process.

New Deployment

**Review** [View CLI Command](#)

**Overview**  
FLOW DEFINITION: kafka\_to\_lakehouse v.1  
ENVIRONMENT DEPLOYING TO: ssa-hol  
DEPLOYMENT NAME: user050

**NiFi Configuration**  
NIIFI 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. The blue box indicates that the Flow deployment process has been started. By clicking on the button **Load More** you will be able to see the different stages of the deployment. After about 60 to 90 seconds approximately, the last event should be *Deployment Successful*.

The screenshot shows the Cloudera DataFlow interface. On the left is a dark sidebar with navigation links: Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and a user profile for Test50 User50. The main area is titled 'Dashboard' and includes filters for 'STATUS All - 15' and 'ENVIRONMENTS All - 1'. A table lists flows by status and name, with one entry for 'user050' currently 'Deploying'. To the right, a detailed view for 'user050' shows a red box highlighting a message: 'Deployment Initiated' with the note 'Initiated deployment of [user050]'. Below this are tabs for KPIs, System Metrics, and Alerts (which is selected). It also shows 'Active Alerts' (none), 'Event History' (none), and a log entry for 'Deployment Initiated' at '2023-05-21 00:09 CEST'. A 'Load More' button is visible.

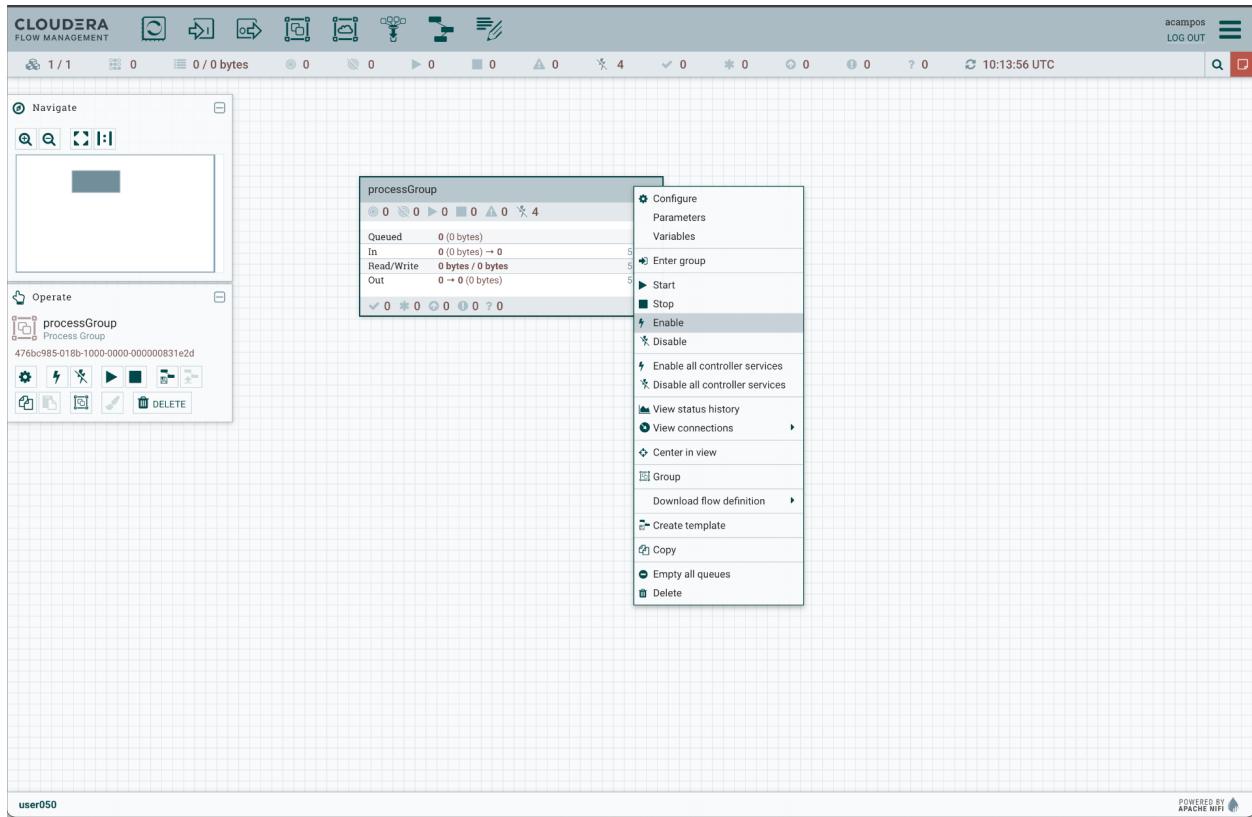
12. Once the deployment is finished, click on **Manage Deployment** to see the details of the recently deployed Flow.

The screenshot shows the Cloudera DataFlow interface. On the left sidebar, there are links for Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, and Environments. The main area is titled 'Dashboard' and shows a table with one row for 'user050'. The row details: Status is 'Deploying', Name is 'user050', and Environment is 'ssa-hol'. To the right of the table, there's a detailed view for 'user050' under 'user050'. It includes tabs for KPIs, System Metrics, and Alerts (which is selected). Below the tabs, it says 'No alerts to display.' and 'Event History'. Under 'Event History', there are several log entries with icons and timestamps. A red box highlights the 'Manage Deployment' button at the top right of the detailed view.

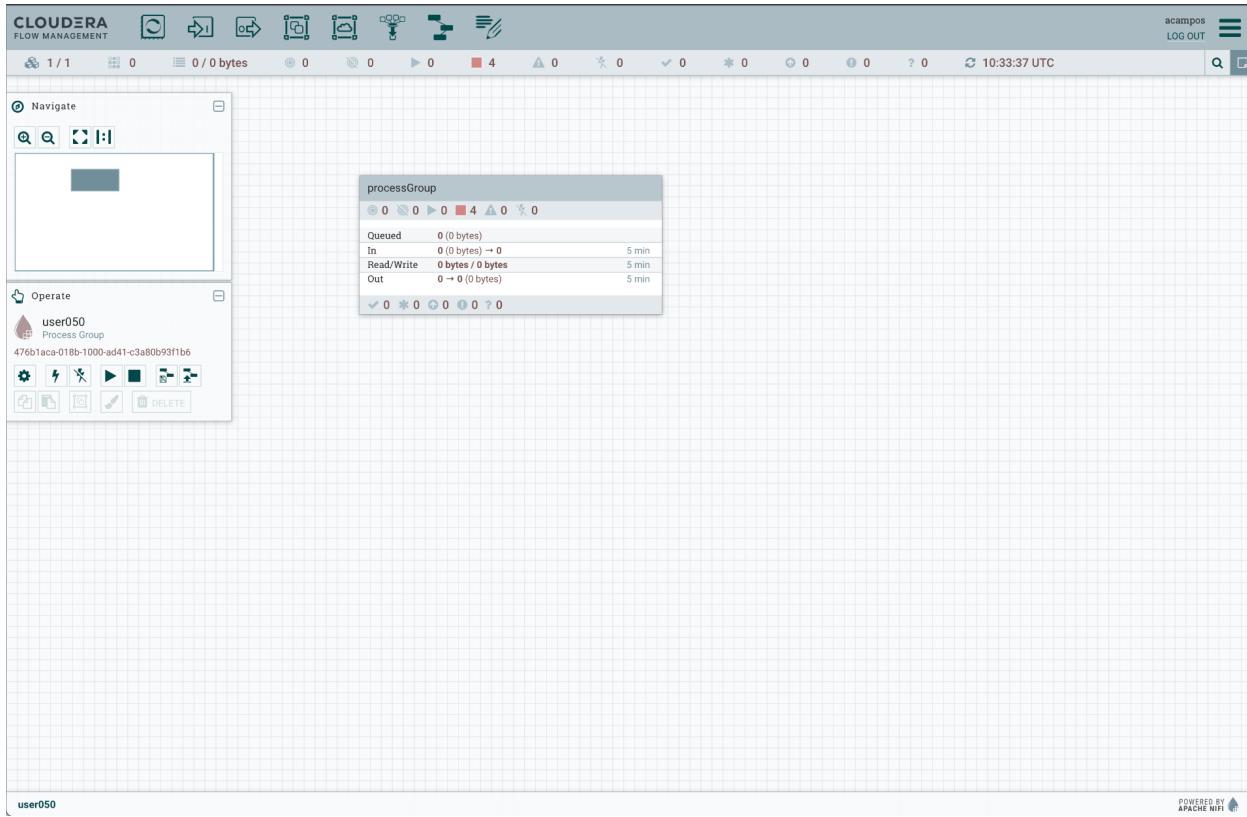
13. In this window you will see the Flow information displayed. It is time to execute the application processes from the graphical Flow Management interface. Click on **Actions -> View in NiFi**, to open Cloudera Flow Management canvas in a new window/tab.

The screenshot shows the 'Deployment Manager' page for deployment 'user050'. The left sidebar has the same navigation as the previous screen. The main content area shows deployment details: STATUS is 'Suspended', NODE COUNT is 1, and ENVIRONMENT is 'aws ssa-hol'. It also lists DEPLOYMENT NAME 'user050', FLOW DEFINITION 'kafka\_to\_lakehouse V1', and various timestamps for creation, last update, and runtime version. On the right, there's a 'Actions' dropdown menu with options: 'View in NiFi', 'Start flow', 'Change NiFi Runtime Version', 'Restart Deployment', and 'Terminate'. Below the deployment details, there's a 'Deployment Settings' section with tabs for KPIs and Alerts, Sizing and Scaling, Parameters, and NiFi Configuration. At the bottom, there's a 'Key Performance Indicators' section with a note to set up KPIs and a 'Learn more' link, followed by a placeholder for KPIs with a 'Add New KPI' button. At the very bottom, there are buttons for 'Discard Changes', 'Apply Changes', and 'Update Deployment CLI Command'.

14. In the new window you should be able to see the Flow Management canvas with one process group (a box) titled **processGroup**. You first need to enable the process group. Right-click on the Process Group and then select the option **Enable**.



15. Then double click on the Process Group to open it.



16. When opening the Process Group, you should be able to see the Processors that compose the Flow application. To summarize, there are four Processors:

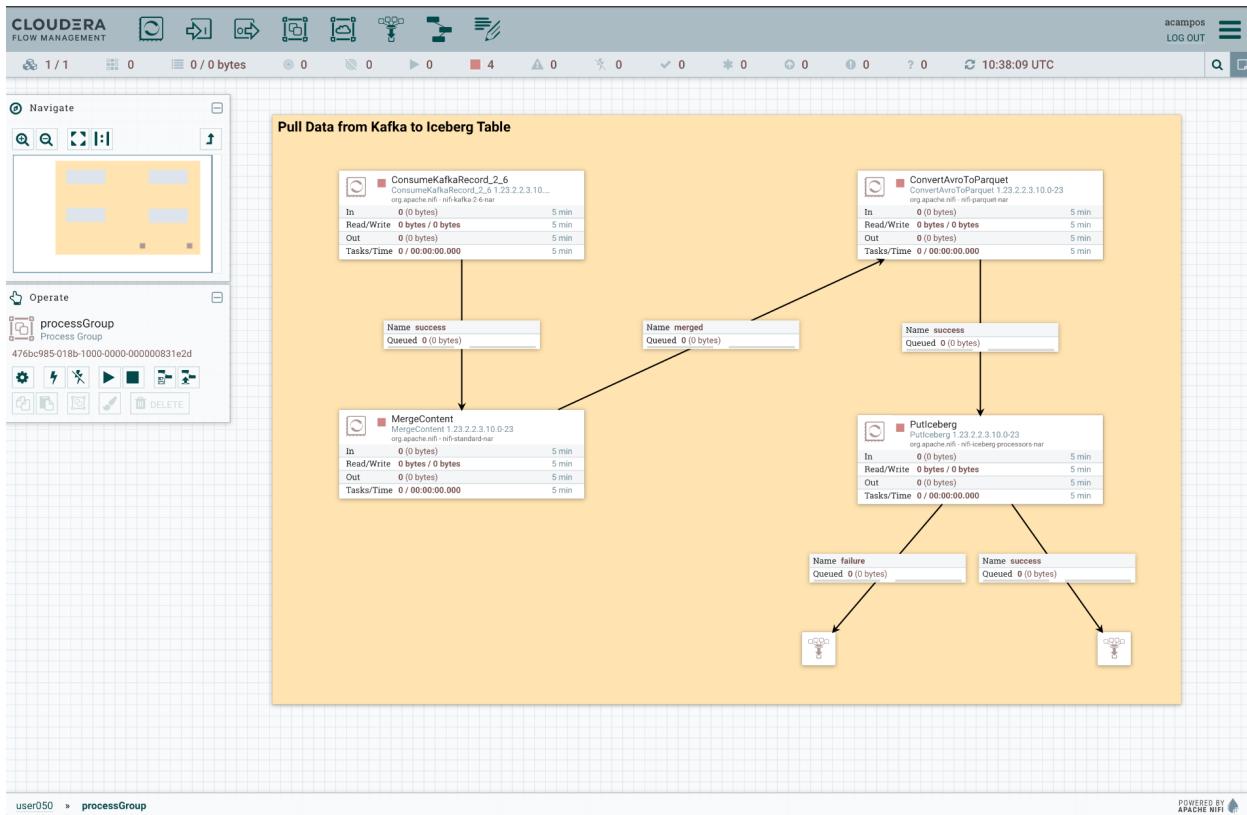
**ConsumeKafkaRecord**, processor to consume data from the Kafka topic, reading the data in JSON format and outputting in AVRO format.

**MergeContent**, to group the flow files and streamline the data flow.

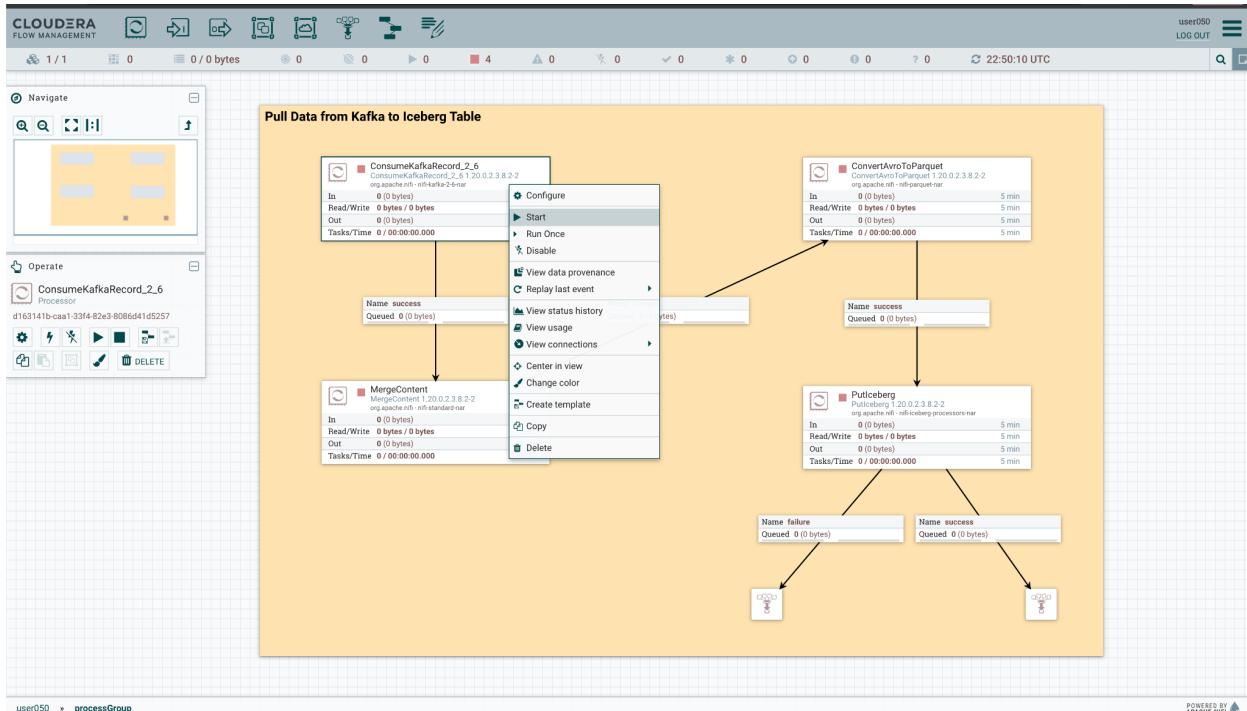
**ConvertAvroToParquet**, conversion needed to store the data in PARQUET format.

**PutIceberg**, to insert the data into the table in the Lakehouse. The destination table is called *telco\_kafka\_iceberg*, and each user has an assigned database (*user\_id* is the name of the database).

As you can see, the Processors are not started, they are paused.

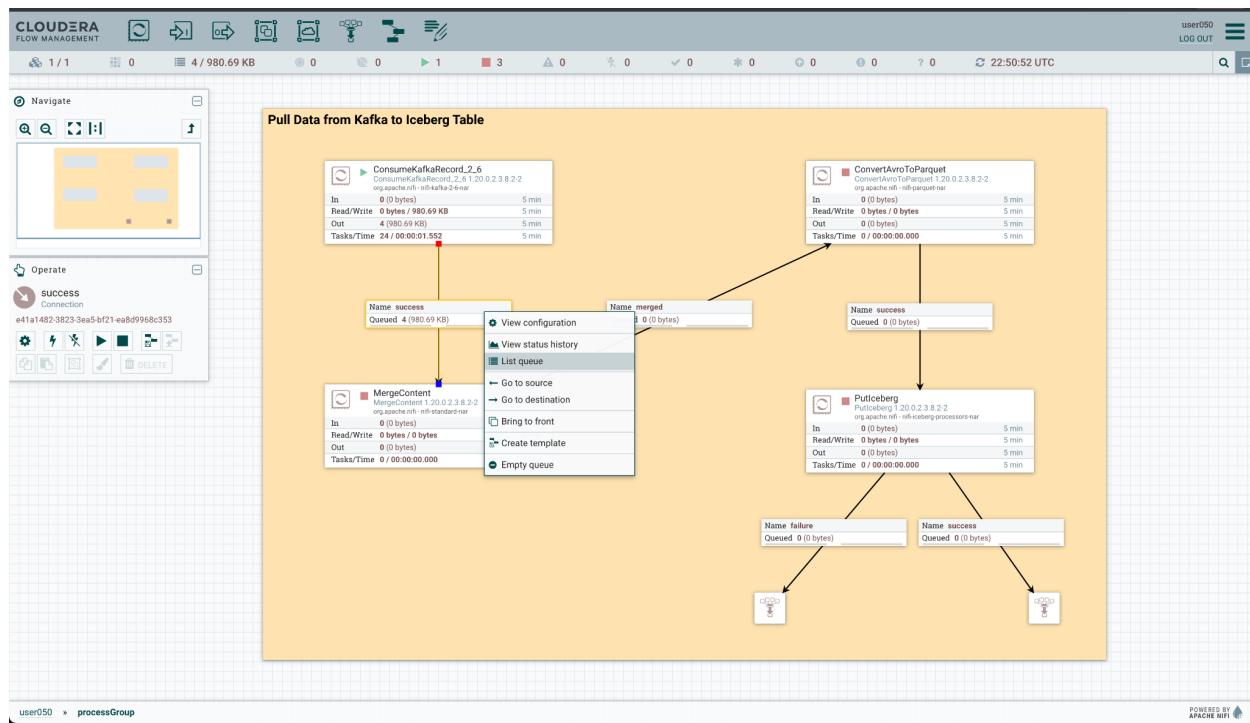


17. Now you are ready to initiate the pipeline. Start with **ConsumeKafkaRecord**, by right-clicking on it, and then clicking on **Start**. This will start consuming the Kafka topic data.



18. Flow Management allows us to see and access data in motion during the execution of the data flow. Between Processors **ConsumeKafkaRecord** (just started) and **MergeContent**, there is a connection. This connection is what joins the Processors and transmits data from one to the other.

To check how much data is queued on this connection, refresh the counter by pressing the Ctrl+R (Windows) or Command+R (Mac) combination on the keyboard. This will allow the current metrics of the entire data stream to be updated. At some point there should be a number next to the legend **Queued** in the connection between **ConsumeKafkaRecord** and **MergeContent**. To see the queued data, right click on the connection and click on the option **List Queue**, opening a popup window.



19. The next popup window lists the queued data. Click on the information icon (i) that appears on the left side to view the events.

The screenshot shows the Apache Nifi User interface with a table titled "Displaying 4 of 4 (980.69 KB)". The table has columns: Position, UUID, Filename, File Size, Queued Duration, Lineage Duration, Penalized, and Node. The data is as follows:

Position	UUID	Filename	File Size	Queued Duration	Lineage Duration	Penalized	Node
1	2055d337-695f-4c6d-8203-3ece27a62d...	2055d337-695f-4c6d-8203-3ece27a62d...	278.24 KB	00:00:12.787	00:00:13.068	No	dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.c...
2	510c8074-9798-4199-a228-ad7894ac9...	510c8074-9798-4199-a228-ad7894ac9...	283.60 KB	00:00:11.664	00:00:11.733	No	dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.c...
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.c...
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.c...

Below the table, a message says "The source of this queue is currently running. This listing may no longer be accurate." A note at the bottom left says "Last updated: 22:50:59 UTC". The URL in the address bar is "user050/processGroup".

20. Once the FlowFile detail window appears, click on the button **VIEW** to open the content of consumed events.

The screenshot shows the Apache Nifi User interface with a modal dialog titled "FlowFile". The dialog has tabs "DETAILS" and "ATTRIBUTES". The "DETAILS" tab is selected, showing the following details for the FlowFile:

FlowFile Details	Content Claim
UUID 2055d337-695f-4c6d-8203-3ece27a62dee	Container default
Filename 2055d337-695f-4c6d-8203-3ece27a62dee	Section 1
File Size 278.24 KB	Identifier 1684623047700-1
Queue Position No value set	Offset 0
Queued Duration 00:00:19.534	Size 278.24 KB
Lineage Duration 00:00:19.815	<a href="#">DOWNLOAD</a> <a href="#">VIEW</a>
Penalized No	
Node Address dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443	

At the bottom right of the dialog is an "OK" button. The URL in the address bar is "user050/processGroup".

21. The new window that opens shows the data of the FlowFile content. Being in AVRO format, it is not fully readable. A deserializer must be selected to correctly display the data. For this, in the upper left, select the option **formatted** from the menu **View as**.

22. Now you can display the data correctly. Notice that the fields or attributes indicated at the beginning of the workshop appear. You can close that FlowFile window and the popups, returning to the canvas with the four Processors.

```

1  [
2   {
3     "multiplelines": "No phone service",
4     "paperlessbilling": "Yes",
5     "gender": "F",
6     "onlinesecurity": "No",
7     "internetservice": "DSL",
8     "contract": "1",
9     "churn": "No",
10    "seniorcitizen": "0",
11    "deviceprotection": "No",
12    "techsupport": "No",
13    "streamingmovies": "No",
14    "totalcharges": "29.85",
15    "monthlycharges": "29.85",
16    "customerid": "7590-VVWBG",
17    "dependents": "0",
18    "onlinebackup": "Yes",
19    "phoneservice": "No",
20    "streamingtvtv": "No",
21    "paymentmethod": "Electronic check"
22  },
23  {
24   {
25     "multiplelines": "No",
26     "paperlessbilling": "No",
27     "gender": "M",
28     "onlinesecurity": "Yes",
29     "internetservice": "DSL",
30     "contract": "2",
31     "churn": "No",
32     "seniorcitizen": "0",
33     "deviceprotection": "Yes",
34     "streamingtvtv": "No",
35     "streamingmovies": "No",
36     "totalcharges": "1889.5",
37     "monthlycharges": "56.95",
38     "customerid": "5575-QNVD8",
39     "dependents": "0",
40     "onlinebackup": "No",
41     "phoneservice": "Yes",
42     "streamingtv": "Yes",
43     "paymentmethod": "Mailed check"
44  },
45  {
46   {
47     "multiplelines": "No",
48     "paperlessbilling": "Yes",
49     "gender": "M",
50     "onlinesecurity": "Yes",
51     "internetservice": "DSL",
52     "contract": "No",
53     "churn": "Yes",
54     "seniorcitizen": "0",
55     "deviceprotection": "No",
56     "streamingtvtv": "No",
57     "streamingmovies": "No",
58     "totalcharges": "108.15",
59     "partner": "N",
60     "dependents": "1",
61     "customerid": "3668-OPV8K",
62     "dependents": "0",
63     "phoneservice": "Yes",
64     "streamingtv": "Yes",
65     "tenure": "2",
66     "paymentmethod": "Mailed check"
67  }
68 ]

```

23. Continue running each of the Processors in order: **MergeContent**, after **ConvertAvroToParquet** and finally **PutIceberg**. Remember that you can refresh the flow counters with the combination Control+R or Command+R.

If the previous steps were executed correctly, the connection of the Processor **PutIceberg** to a funnel should be of type **success**.

