

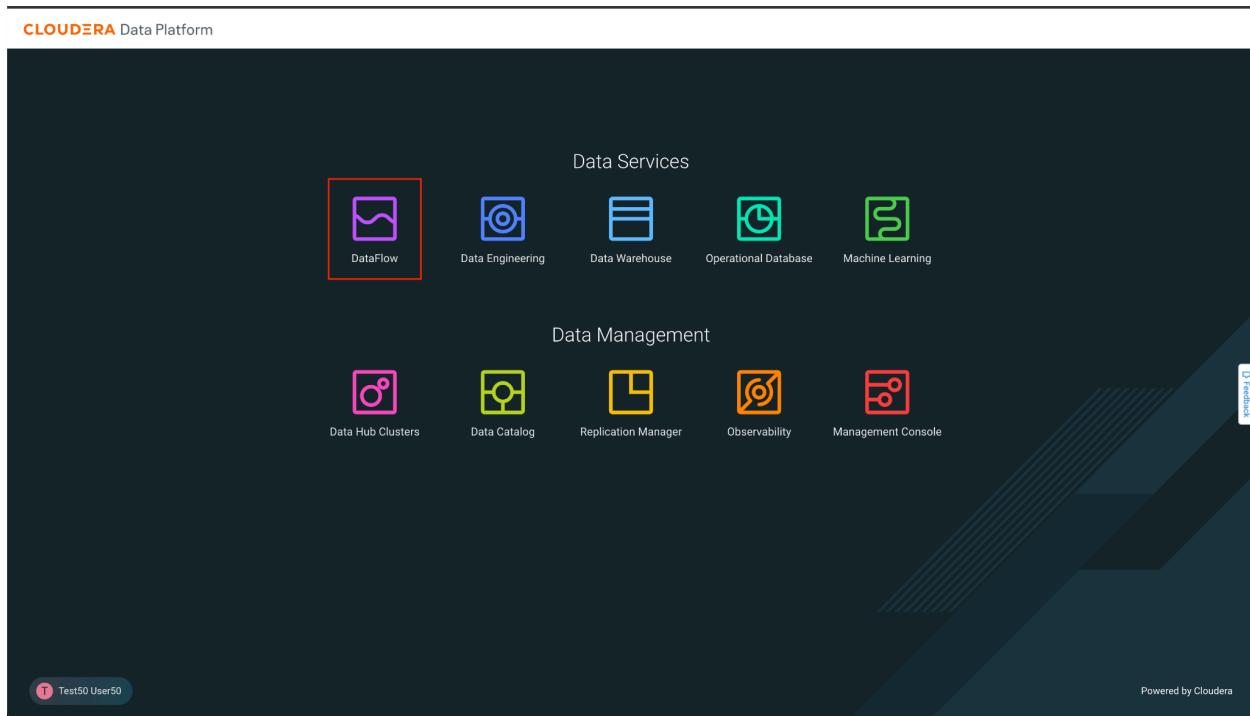
# Data Lifecycle CDP Public Cloud

## 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:



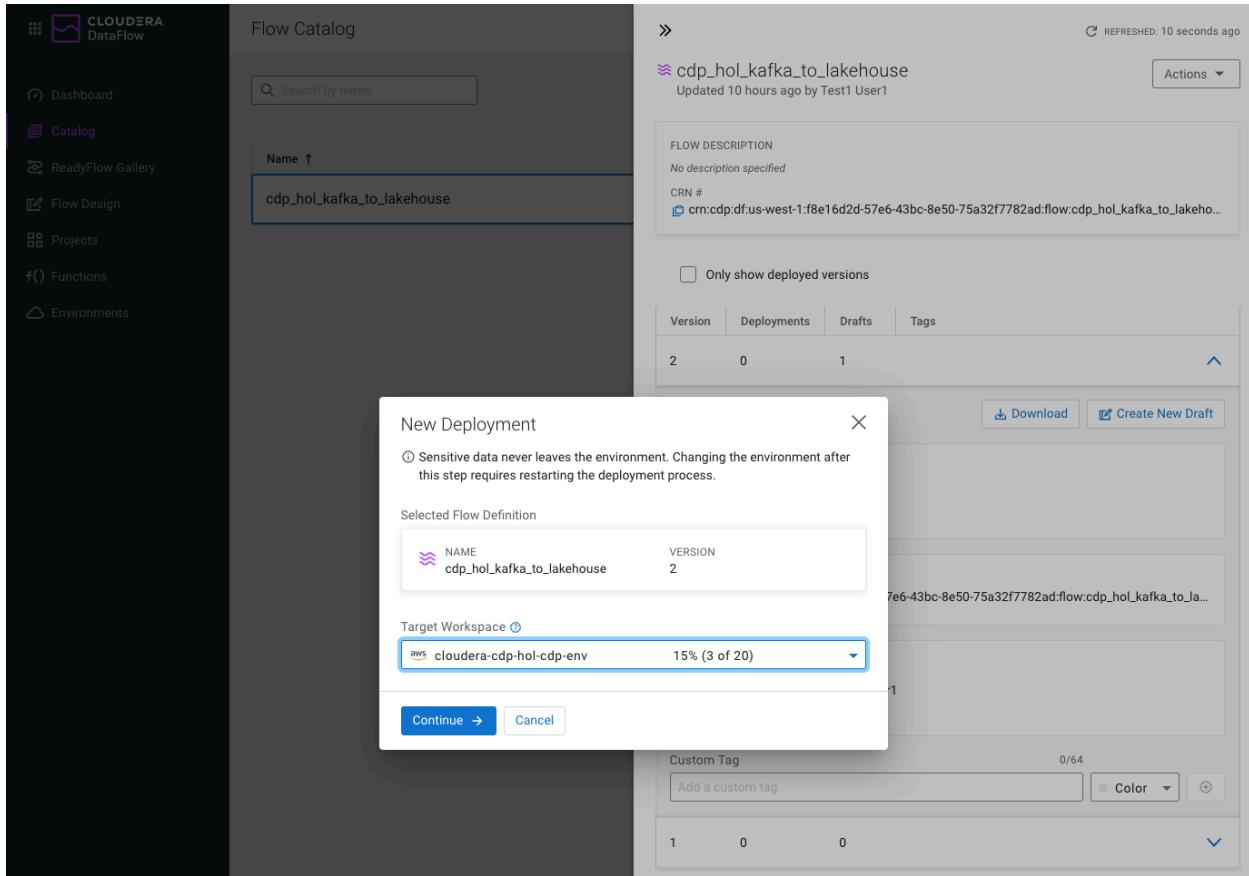
2. Once in DataFlow, click on the option **Catalog** from the left menu. The data ingestion application templates are listed here. For the purpose of this workshop, we have created and published a template that allows you to read Kafka topic data and ingest/store it in the Lakehouse provided by CDP Public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

The screenshot shows the Cloudera DataFlow interface. On the left is a dark sidebar with the Cloudera logo and the text "CLOUDERA DataFlow". Below the logo are several menu items: Dashboard, Catalog (which is selected and highlighted in purple), ReadyFlow Gallery, Flow Design, Projects, Functions, and Environments. The main area is titled "Flow Catalog" and contains a search bar with the placeholder "Search by name" and a blue button labeled "Import Flow Definition". A message at the top right says "REFRESHED: 19 seconds ago". Below the search bar is a table header with columns: Name ↑, Type, Versions, and Last Updated. A single row is visible in the table, showing "cdp\_hol\_kafka\_to\_lakehouse" as the Name, "Custom Flow Definition" as the Type, "2" as the Versions count, and "10 hours ago" as the Last Updated time. At the bottom of the table 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.

The screenshot shows the Cloudera DataFlow interface. On the left is a dark sidebar with navigation links: Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Projects, Functions, and Environments. The main area is titled "Flow Catalog" and contains a search bar. A list of flows is shown, with "cdp\_hol\_kafka\_to\_lakehouse" selected. The right side displays detailed information about this flow, including its name, last refresh time (6 seconds ago), and a "Actions" dropdown. Below this is the "FLOW DESCRIPTION" section, which notes "No description specified" and provides the CRN: crn:cdp:df:us-west-1:f8e16d2d-57e6-43bc-8e50-75a32f7782ad:flow:cdp\_hol\_kafka\_to\_lakeho...". There is also a checkbox for "Only show deployed versions". A table shows deployment counts: Version 2 has 0 Deployments and 1 Draft. Below this are buttons for "Deploy" (with a progress bar), "Download", and "Create New Draft". The "ASSOCIATED DRAFTS (1)" section lists "aws cloudera-cdp-hol-cdp-env" with "cdp-hol-kafka-to-lakehouse" as a draft. The "CRN #" section shows the same CRN as above. The "CREATED" section shows the date "2024-04-24 00:32 PDT" and the user "Test1 User1". The "Custom Tag" section allows adding a custom tag with a maximum length of 64 characters, a color picker, and a plus sign for new tags. Deployment counts at the bottom are 1, 0, and 0.

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 **cloudera-cdp-hol-cdp-env**. 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 **Deployment Name**, **Target Project**, and click **Next**.

For the purposes of this workshop, please name the Flow starting with your assigned username. For example, **user000**

Select **workshop** or **unassigned** for the project, which is a way to organize your flows.

## New Deployment

1 Overview

2 NiFi Configuration

3 Parameters

4 Sizing & Scaling

5 Key Performance Indicators

6 Review

### Overview

Deployment Name

user001-kafka-to-lakehouse

✓ Deployment name is valid

Selected Flow Definition

NAME	VERSION
cdp_hol_kafka_to_lakehouse	2

Target Environment

aws cloudera-cdp-hol-cdp-env

Target Project ⓘ

cdp-hol-workshop

Filter by name

Unassigned

cdp-hol-source

cdp-hol-workshop

Cancel

Next →

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

## New Deployment

The screenshot shows the 'NiFi Configuration' step of a deployment wizard. On the left, a vertical navigation bar lists steps 1 through 6: Overview, NiFi Configuration (selected), Parameters, Sizing & Scaling, Key Performance Indicators, and Review. The main area contains configuration options:

- NiFi Runtime Version:** CURRENT VERSION Latest Version (1.24.0.2.3.12.2-1). Includes a 'Change Version' link.
- Autostart Behavior:** A checked checkbox for 'Automatically start flow upon successful deployment'.
- Inbound Connections:** An unchecked checkbox for 'Allow NiFi to receive data'.
- Custom NAR Configuration:** An unchecked checkbox for 'This flow deployment uses custom NARs'.

On the right, a sidebar titled 'Overview' displays deployment details:

- FLOW DEFINITION: cdp\_hol\_kafka\_to\_lakehouse v.2
- ENVIRONMENT DEPLOYING TO: cloudera-cdp-hol-cdp-env
- PROJECT ASSIGNING TO: cdp-hol-workshop
- DEPLOYMENT NAME: user001-kafka-to-lakehouse

At the bottom are buttons for 'Cancel', 'Previous', and 'Next →'.

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

**workload\_password:** Enter the Workload Password shared at the beginning of the workshop.

**workload\_user:** Enter the assigned <user id> (example. user-050)

**Database:** Enter <user id>

**Kafka Brokers:**

cdp-hol-smm-corebroker2.cloudera.z30z-14kp.cloudera.site:9093,cdp-hol-smm-corebroker1.cloudera.z30z-14kp.cloudera.site:9093,cdp-hol-smm-corebroker0.cloudera.z30z-14kp.cloudera.site:9093

**Kafka Consumer group Id:** Enter <user id>-group (eg. user050-group)

**Kafka Topic:** telco\_data

NOTE: for the purposes of the workshop, your user (e.g. user050) is also the name of the **database** where you will store the data (which has already been created for you), and the name of the **Kafka Consumer Group ID** for reading messages.

For the purposes of this workshop, the remaining values were filled out for you and don't need to change.

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

## New Deployment

Overview

NiFi Configuration

Parameters

Sizing & Scaling

Key Performance Indicators

Review

cdp-hol-kafka-to-lakehouse (7)

CDP Workload User Password  
Enter parameter values. 0/100K

CDP Workload Username  
user001 7/100K

CDPEnvironment ⓘ

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

Select File  
Drop file or browse 7/100K

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  
user001 7/100K

Kafka Brokers  
cdp-hol-smm-corebroker2.cloudera.z30z-14kp.cloudera.site:9093,cdp-hol-smm-corebroker1.cloudera.z30z-14kp.cloudera.site:9093,cdp-hol-smm-corebroker0.cloudera.z30z-14kp.cloudera.site:9093 185/100K

Kafka Consumer Group Id  
user001-group2 14/100K

Kafka Topic  
telco\_data 10/100K

Cancel ← Previous Next →

Overview

FLOW DEFINITION  
cdp\_hol\_kafka\_to\_lakehouse v.2

ENVIRONMENT DEPLOYING TO  
cloudera-cdp-hol-cdp-env

PROJECT ASSIGNING TO  
cdp-hol-workshop

DEPLOYMENT NAME  
user001-kafka-to-lakehouse

NiFi Configuration

NIFI RUNTIME VERSION  
Latest Version (1.24.0.2.3.12.2-1)

AUTO-START FLOW  
Yes

INBOUND CONNECTIONS  
No

CUSTOM NAR CONFIGURATION  
No

8. There is no need to configure auto-scaling parameters. Click **Next**.

New Deployment

**Sizing & Scaling**  
Select the NiFi node size and the number of nodes provisioned for your flow.

**NiFi Node Sizing**

<input checked="" type="radio"/> Extra Small	<input type="radio"/> Small	<input type="radio"/> Medium	<input type="radio"/> Large
2 vCores Per Node 4 GB Per Node	3 vCores Per Node 6 GB Per Node	6 vCores Per Node 12 GB Per Node	12 vCores Per Node 24 GB Per Node

**Number of NiFi Nodes**  
Auto Scaling  Disabled

Nodes:  1 to 32

**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.yu1-vbzg.cloudera.site:9093,realtime-ingestion-corebroker1.ssa-hol.yu1-vbzg.cloudera.site:9093,realtime-ingestion-corebroker2.ssa-hol.yu1-vbzg.cloudera.site:9093

**Cancel** **← Previous** **Next →**

9. We are also not going to configure KPIs now. Click **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](#)

**Add New KPI**

**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.yu1-vbzg.cloudera.site:9093,realtime-ingestion-corebroker1.ssa-hol.yu1-vbzg.cloudera.site:9093,realtime-ingestion-corebroker2.ssa-hol.yu1-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

The screenshot shows the 'Review' step of a deployment process. On the left, a sidebar lists steps: Overview, NiFi Configuration, Parameters, Sizing & Scaling, Key Performance Indicators, and Review (selected). The main area displays the 'Review' section with the following details:

- 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:

At the bottom are buttons for **Cancel**, **← Previous**, and **Deploy**.

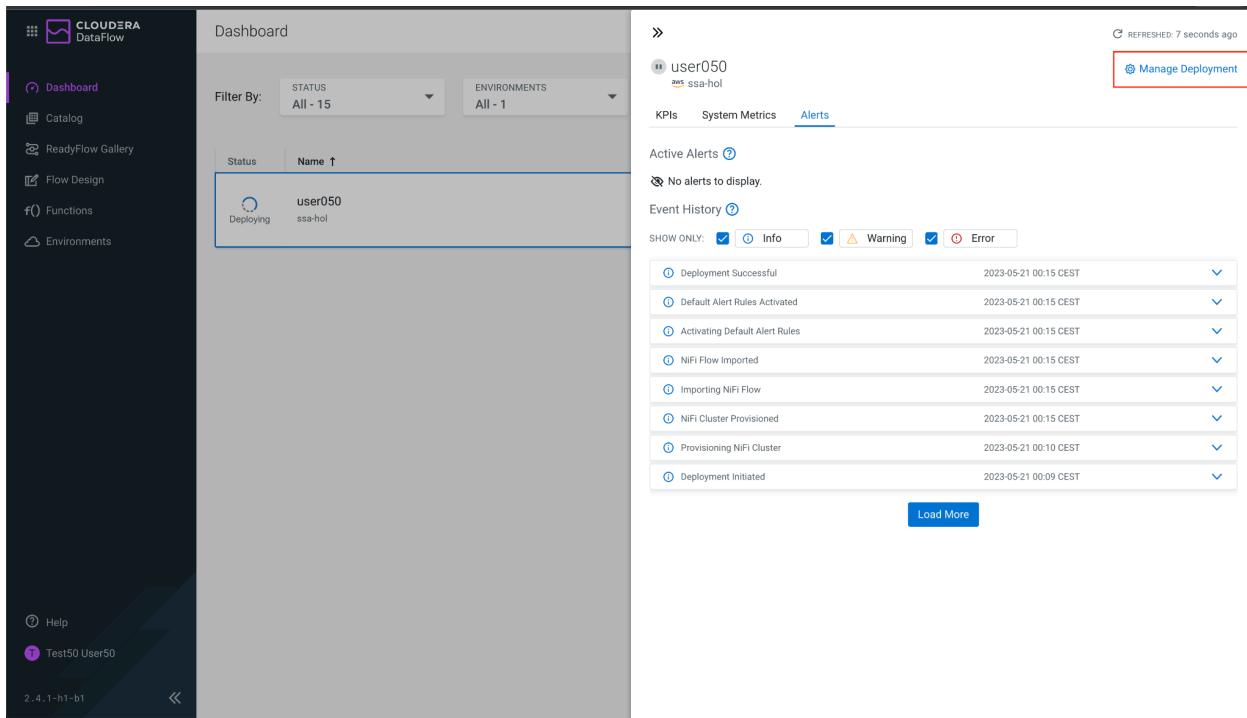
11. The blue box indicates that the Flow deployment process has been started. By clicking on the button **Load More** you will be able to see the different stages of the deployment. After about 60 to 90 seconds approximately, the last event should be *Deployment Successful*.

The screenshot shows the Cloudera DataFlow dashboard. On the left, a sidebar includes links for DataFlow, Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and a Test50 User50 entry. The main dashboard area shows a table with columns: Status and Name (sorted by Name). One row is highlighted with a blue border, showing 'Deploying' status for 'user050'.

In the top right corner, there is a message box with a red border containing the text: **Deployment Initiated** Initiated deployment of [user050].

Below the message box, the dashboard displays Active Alerts (No alerts), Event History (Deployment Initiated at 2023-05-21 00:09 CEST), and a 'Load More' button.

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 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 shows a table with one row: 'user050' (Status: Deploying). To the right of the table is a detailed view for 'user050' on the 'aws ssa-hol' environment. It includes tabs for KPIs, System Metrics, and Alerts (which is selected). Below the tabs is a section for 'Active Alerts' which says 'No alerts to display.' Under 'Event History', there is a table of events:

Event	Date
Deployment Successful	2023-05-21 00:15 CEST
Default Alert Rules Activated	2023-05-21 00:15 CEST
Activating Default Alert Rules	2023-05-21 00:15 CEST
NiFi Flow Imported	2023-05-21 00:15 CEST
Importing NiFi Flow	2023-05-21 00:15 CEST
NiFi Cluster Provisioned	2023-05-21 00:15 CEST
Provisioning NiFi Cluster	2023-05-21 00:10 CEST
Deployment Initiated	2023-05-21 00:09 CEST

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.

REFRESHED: 6 seconds ago

**Deployment Manager**

**Deployment Name:** kafka\_to\_lakehouse\_V1

**Created On:** 2023-05-21 00:09 CEST

**Last Updated:** 2023-05-21 00:15 CEST

**Deployed By:** Test50 User50

**CRN #:** crn:cdp:df:us-west-2:  
user050

**Actions**

- View in NiFi
- Start flow
- Change NiFi Runtime Version
- Restart Deployment
- Terminate

**Deployment Settings**

KPIs and Alerts   Sizing and Scaling   Parameters   NiFi Configuration

**Key Performance Indicators**

Add New KPI

**Deployment CLI Command:** > Recreate Deployment CLI Command   > Update Deployment CLI Command

14. Double-click on the Process Group to open it.

acampos LOG OUT

**processGroup**

Count	Value	Unit
Queued	0 (0 bytes)	
In	0 (0 bytes) → 0	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 → 0 (0 bytes)	5 min

**Operate**

user050  
Process Group  
476b1aca-018b-1000-ad41-c3a80b93f1b6

POWERED BY APACHE NIFI

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:

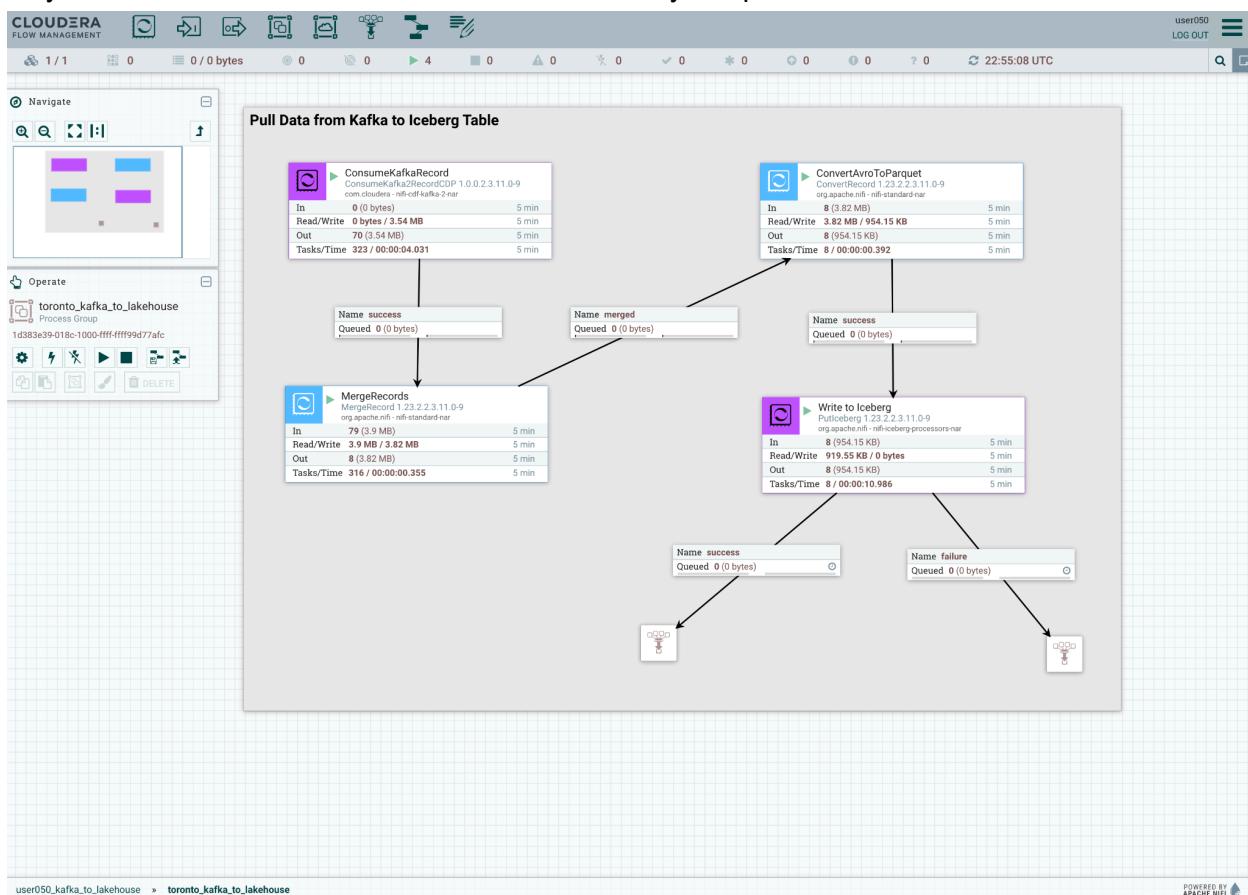
**ConsumeKafkaRecord**, consumes data from the Kafka topic, reading the data in JSON and outputting in AVRO.

**MergeRecords**, 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.

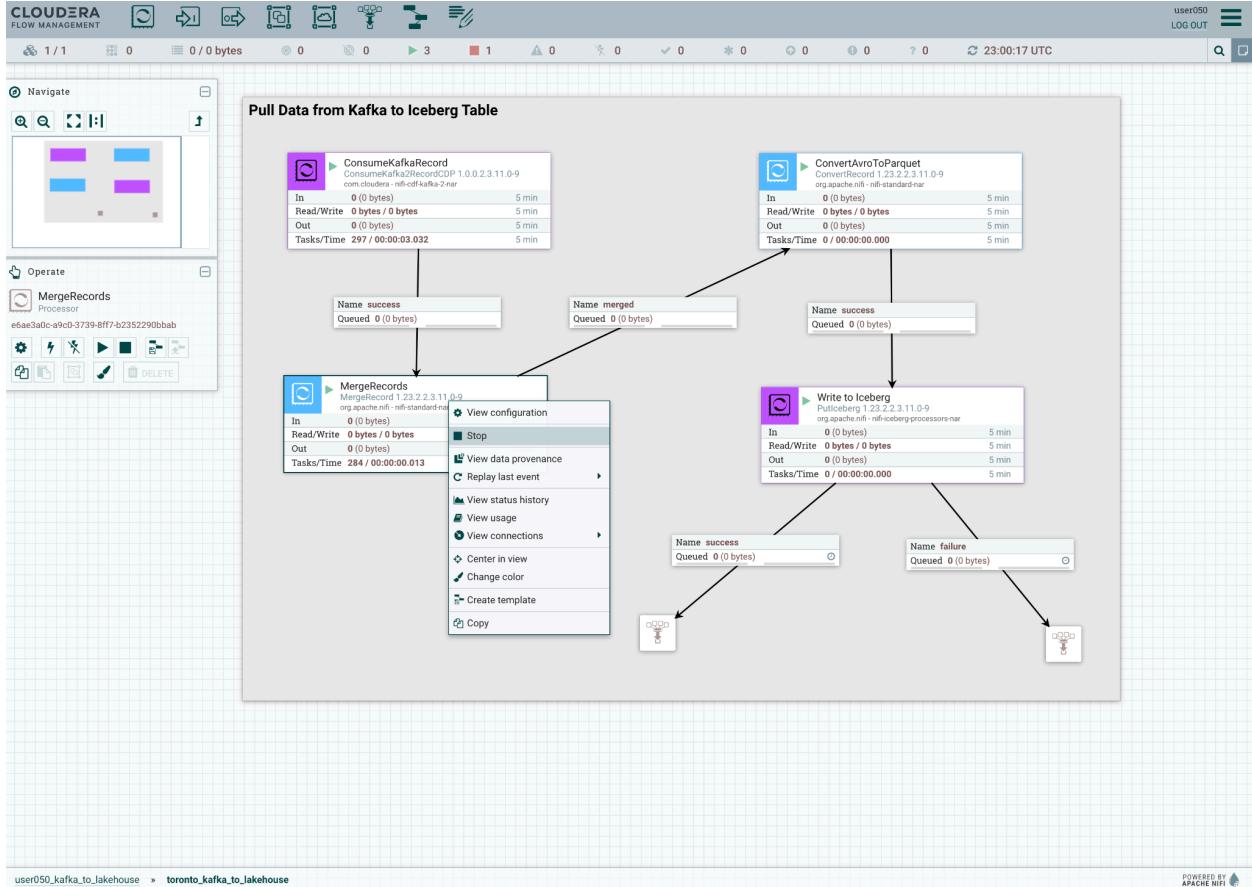


From the Out field in every processor, you can see that data has flowed through in the past 5 minutes. You have already consumed data from Kafka and to Iceberg!

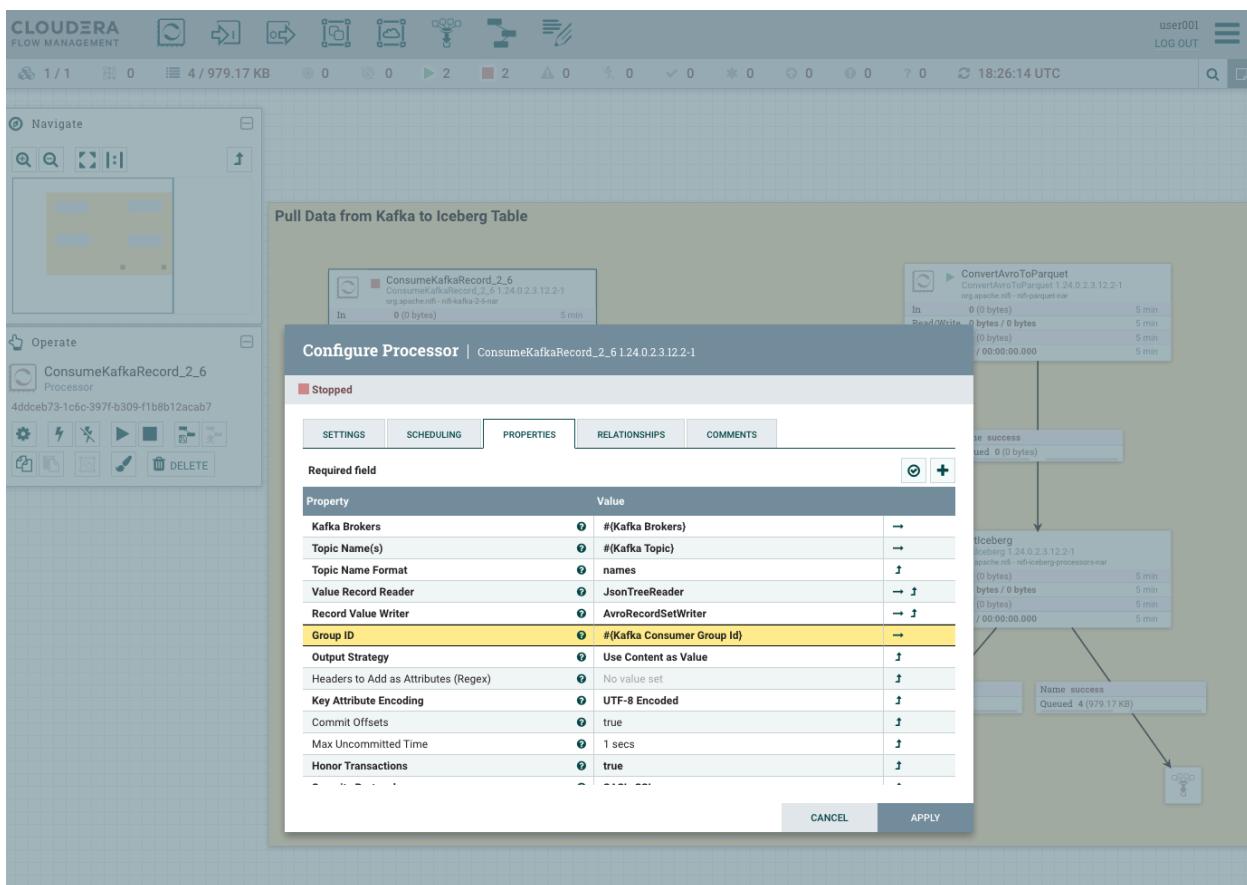
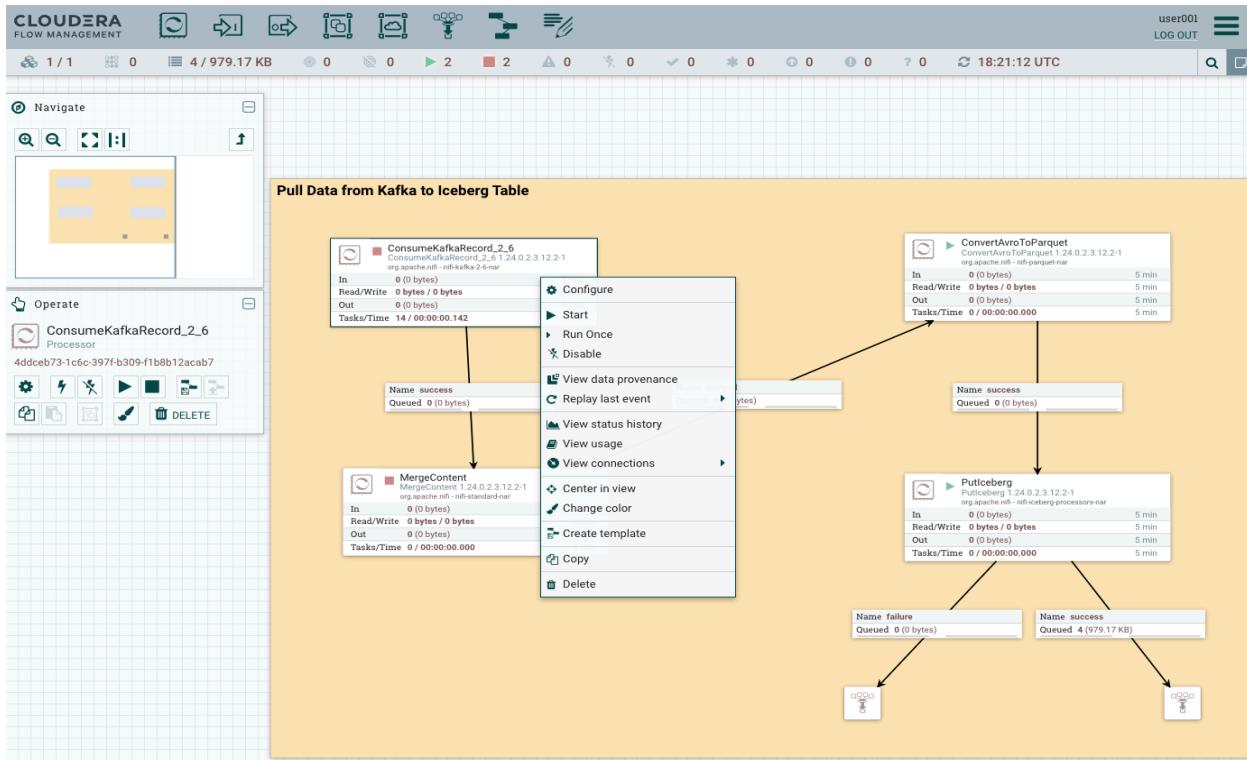
17. Flow Management allows us to see and access data in motion during the execution of the data flow. Between Processors **ConsumeKafkaRecord** (just started) and **MergeRecords**, there is a connection. This connection is what joins the Processors and transmits data from one to the other, and you can check how much data is queued at every step of the process.

Let's see this in action by building up the queue. In this case, existing kafka data already consumed and we want to explore other features of the NiFi processor. We are configuring **ConsumeKafkaProcessor** to pull the data with different kafka consumer groups.

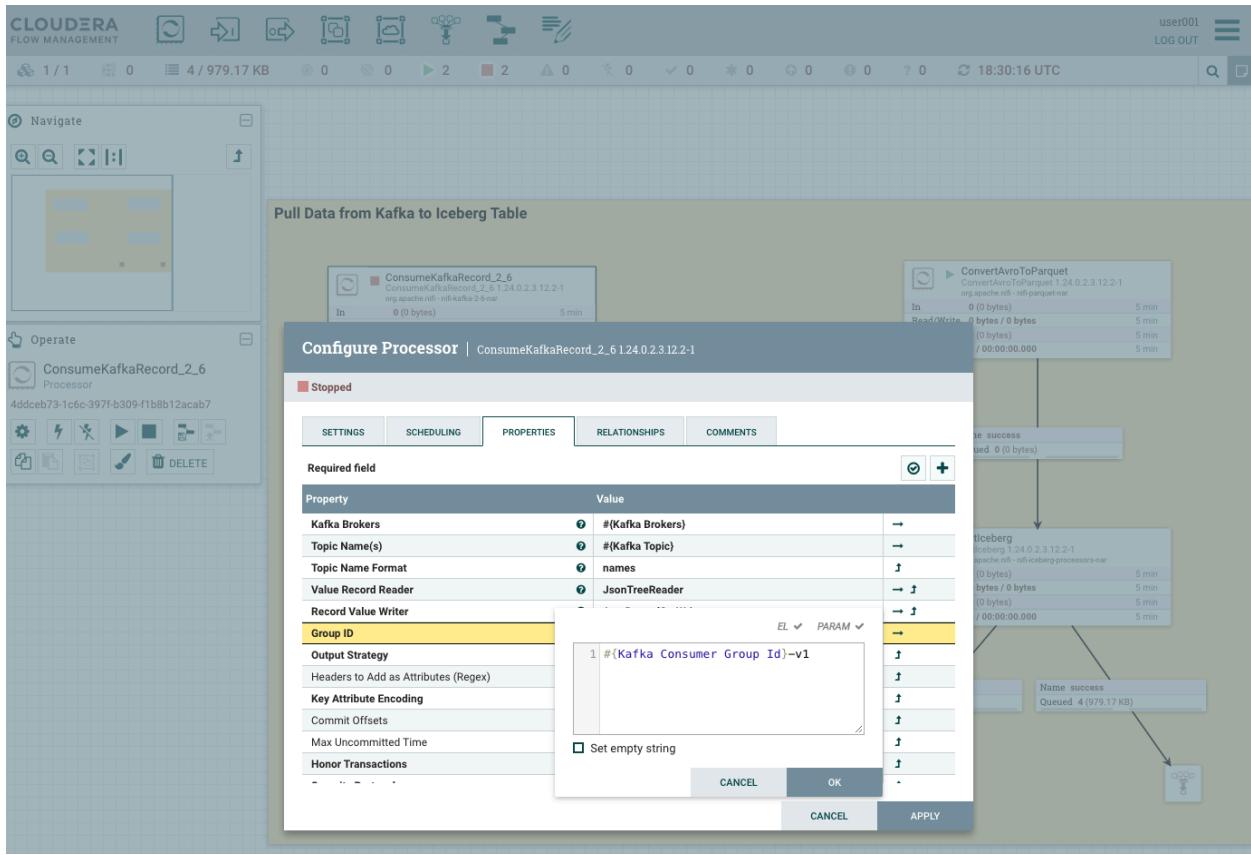
First, right-click on **MergeRecords** processor and click **Stop** and repeat the same for processor **ConsumeKafkaRecord**



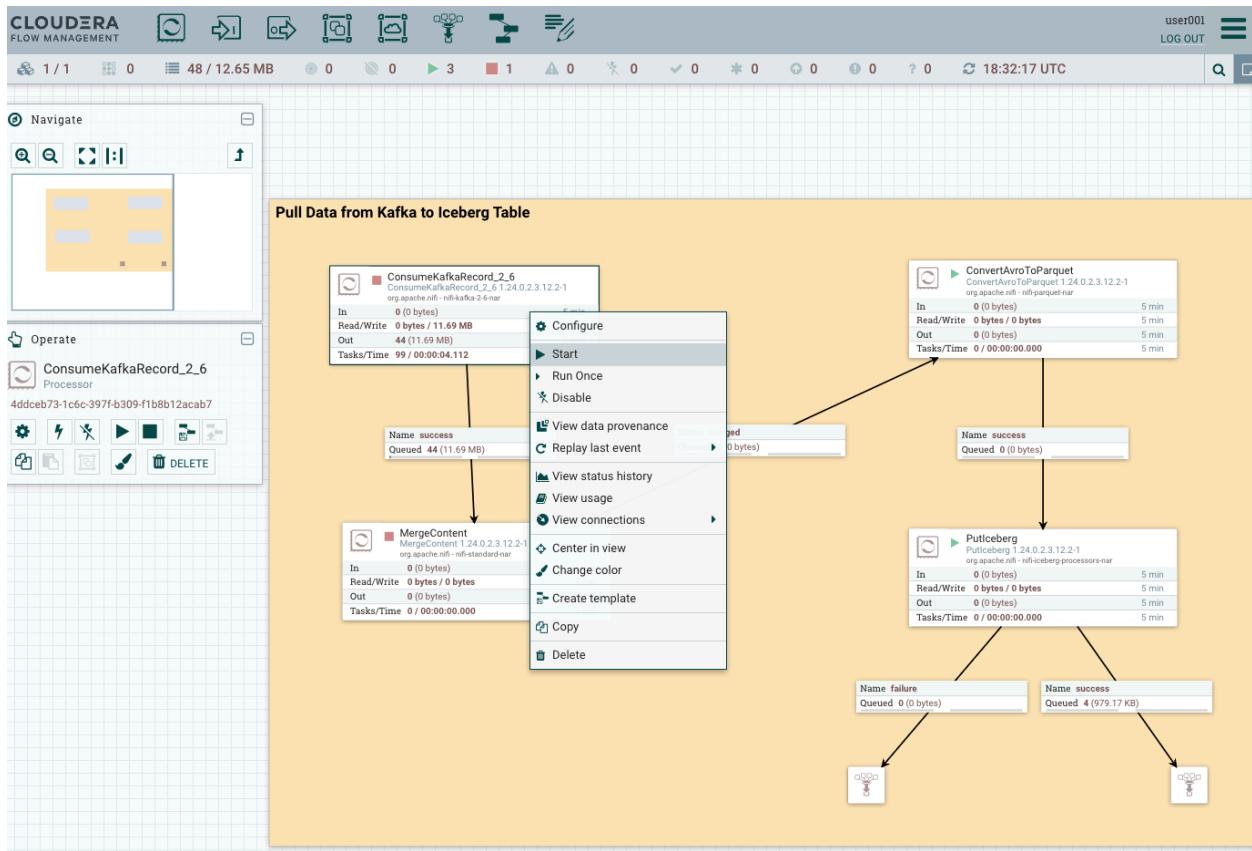
18. right click on ConsumerKafkaRecord processor and click on Configure



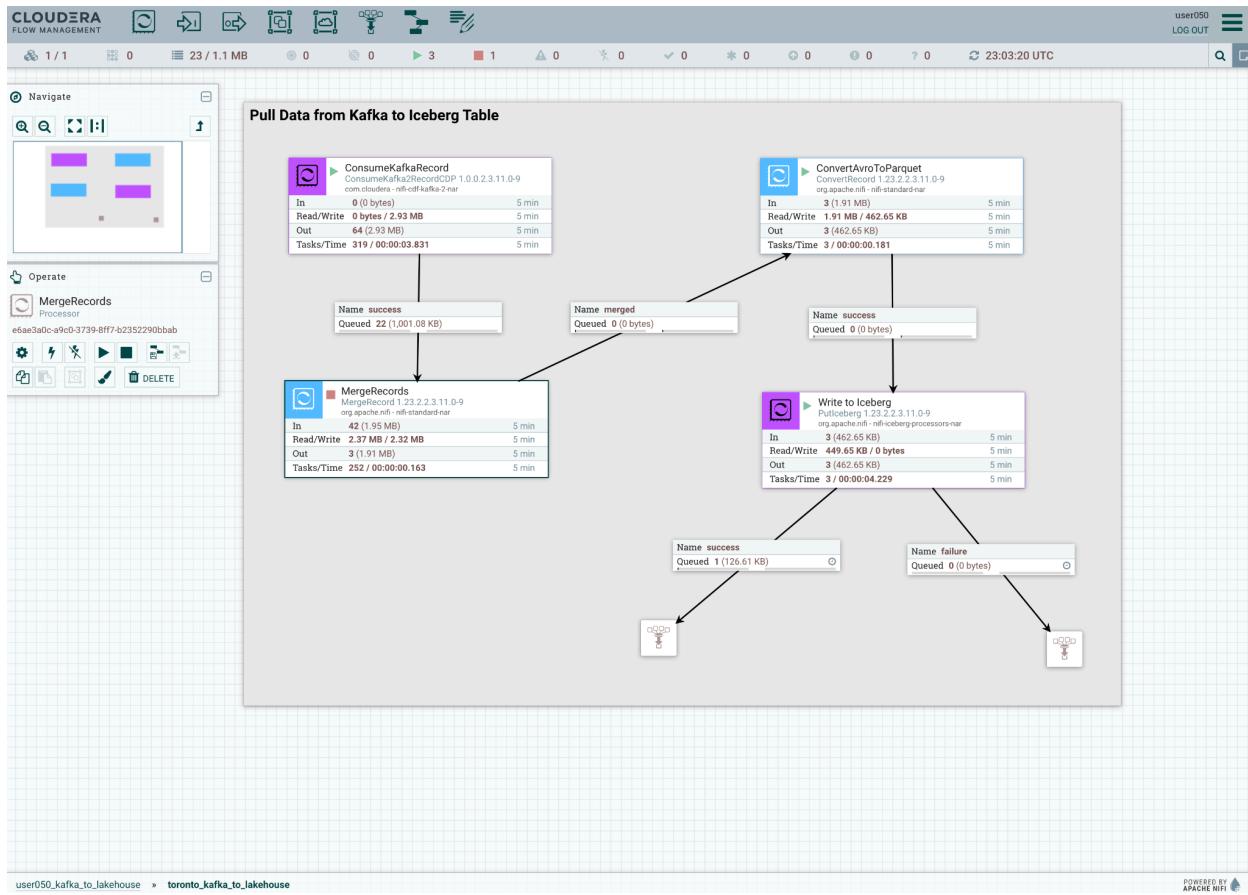
19. Change the consumer group id by Click on group ID value and modify the consumer group id (eg. #{Kafka Consumer Group Id}-v1) and click “OK” and “Apply”



## 20. Start the ConsumerKafkaRecord processor by right clicking and start

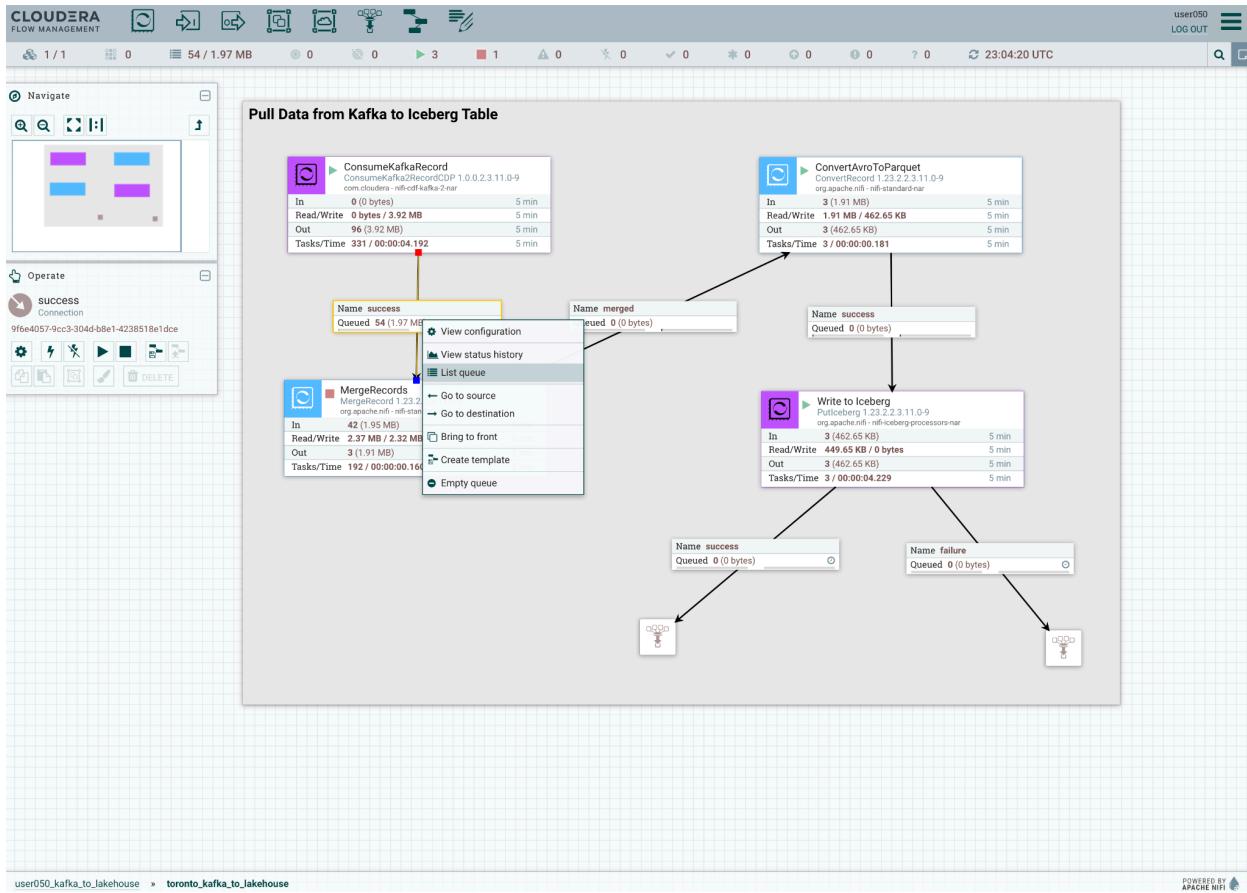


21. You will see data start to queue up in the connector shortly after starting



You can 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 **MergeRecords**. To see the queued data, right-click on the connection and click on the option **List Queue**, opening a popup window.



22. 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 list of FlowFiles in a queue. The title bar says "SUCCESS" and "Displaying 4 of 4 (980.69 KB)". A message at the top right states: "The source of this queue is currently running. This listing may no longer be accurate." Below this is a table with columns: Position, UUID, Filename, File Size, Queued Duration, Lineage Duration, Penalized, and Node. The table contains four rows of data. Row 1: Position 1, UUID 2055d337-695f-4c6d-8203-3ece27a62d..., Filename 2055d337-695f-4c6d-8203-3ece27a62d..., File Size 278.24 KB, Queued Duration 00:00:12.787, Lineage Duration 00:00:13.068, Penalized No, Node dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443. Row 2: Position 2, UUID 510c8074-9798-4199-a228-ad7894ac9..., Filename 510c8074-9798-4199-a228-ad7894ac9..., File Size 283.60 KB, Queued Duration 00:00:11.664, Lineage Duration 00:00:11.733, Penalized No, Node dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443. Row 3: Position 3, UUID cad12e7c-e301-439c-85b3-a53fb0f13a2..., Filename cad12e7c-e301-439c-85b3-a53fb0f13a2..., File Size 285.48 KB, Queued Duration 00:00:11.575, Lineage Duration 00:00:11.647, Penalized No, Node dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443. Row 4: Position 4, UUID 01ee7d33-8e54-4a2b-a39c-a3f965b3cf87, Filename 01ee7d33-8e54-4a2b-a39c-a3f965b3cf87, File Size 133.37 KB, Queued Duration 00:00:11.527, Lineage Duration 00:00:11.567, Penalized No, Node dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443. At the bottom left, it says "Last updated: 22:50:59 UTC". The bottom right shows the Apache Nifi logo.

23. 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 detailed view of a FlowFile. The title bar says "SUCCESS" and "Displaying 4 of 4 (980.69 KB)". A message at the top right states: "The source of this queue is currently running. This listing may no longer be accurate." Below this is a table with columns: Position, UUID, Filename, File Size, Queued Duration, Lineage Duration, Penalized, and Node. The table contains four rows of data. Row 1: Position 1, UUID 2055d337-695f-4c6d-8203-3ece27a62d..., Filename 2055d337-695f-4c6d-8203-3ece27a62d..., File Size 278.24 KB, Queued Duration 00:00:12.787, Lineage Duration 00:00:13.068, Penalized No, Node dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443. Row 2: Position 2, UUID 510c8074-9798-4199-a228-ad7894ac9..., Filename 510c8074-9798-4199-a228-ad7894ac9..., File Size 283.60 KB, Queued Duration 00:00:11.664, Lineage Duration 00:00:11.733, Penalized No, Node dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443. Row 3: Position 3, UUID cad12e7c-e301-439c-85b3-a53fb0f13a2..., Filename cad12e7c-e301-439c-85b3-a53fb0f13a2..., File Size 285.48 KB, Queued Duration 00:00:11.575, Lineage Duration 00:00:11.647, Penalized No, Node dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443. Row 4: Position 4, UUID 01ee7d33-8e54-4a2b-a39c-a3f965b3cf87, Filename 01ee7d33-8e54-4a2b-a39c-a3f965b3cf87, File Size 133.37 KB, Queued Duration 00:00:11.527, Lineage Duration 00:00:11.567, Penalized No, Node dfx-nifi-0.dfx-nifi.dfx-user050-ns.svc.cluster.local:8443. A modal window titled "FlowFile" is displayed, showing "FlowFile Details" and "Attributes". The "DETAILS" tab is selected. It lists various attributes: Content Claim (Container: default, Section: 1), Identifier (1684623047700-1), Offset (0), Size (278.24 KB), and a timestamp (00:00:19.815). There are two buttons at the bottom: "DOWNLOAD" and "VIEW". The "VIEW" button is highlighted with a red border. At the bottom right of the modal is an "OK" button. The bottom left of the modal says "Last updated: 22:50:59 UTC". The bottom right shows the Apache Nifi logo.

24. 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**.



25. 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.

View as: formatted

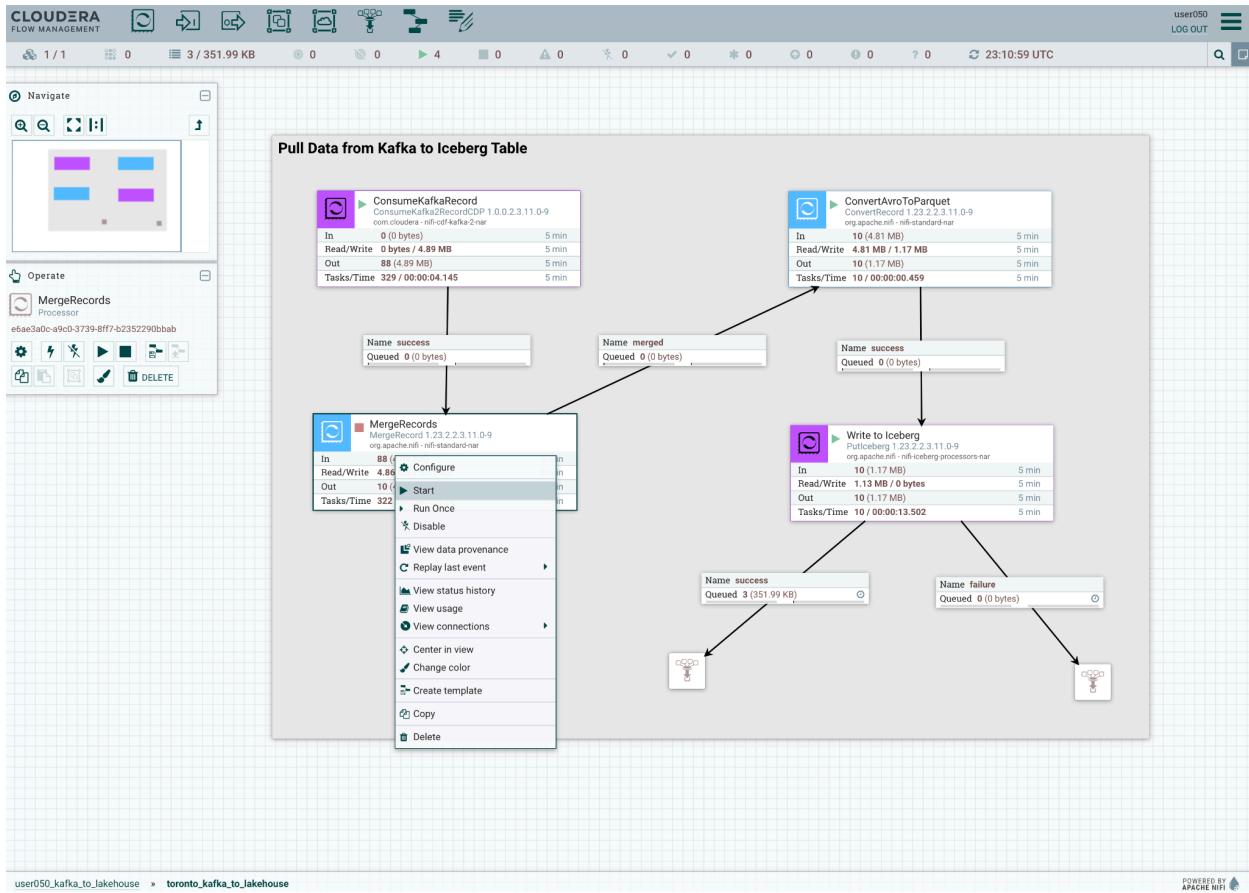
```

1  [
2   {
3     "multiplelines": "No phone service",
4     "paperlessbilling": "Yes",
5     "partner": "Yes",
6     "onlinesecurity": "No",
7     "internetservice": "DSL",
8     "techsupport": "No",
9     "contract": "1",
10    "churn": "Yes",
11    "seniorcitizen": "0",
12    "deviceprotection": "No",
13    "streamingtv": "No",
14    "totalcharges": "29.85",
15    "dependents": "0",
16    "monthlycharges": "29.85",
17    "customerid": "7590-VVWBG",
18    "dependents": "0",
19    "onlinebackup": "Yes",
20    "phoneservice": "No",
21    "streamingmovies": "No",
22    "paymentmethod": "Electronic check"
23  },
24  {
25    "multiplelines": "No",
26    "paperlessbilling": "No",
27    "partner": "Yes",
28    "onlinesecurity": "Yes",
29    "internetservice": "DSL",
30    "techsupport": "Yes",
31    "contract": "2",
32    "churn": "No",
33    "seniorcitizen": "0",
34    "deviceprotection": "Yes",
35    "streamingtv": "No",
36    "totalcharges": "1889.5",
37    "dependents": "0",
38    "monthlycharges": "56.95",
39    "customerid": "5575-QNVD8",
40    "dependents": "0",
41    "onlinebackup": "No",
42    "phoneservice": "Yes",
43    "streamingmovies": "Yes",
44    "paymentmethod": "Mailed check"
45  },
46  {
47    "multiplelines": "No",
48    "paperlessbilling": "Yes",
49    "gender": "M",
50    "partner": "Yes",
51    "onlinesecurity": "Yes",
52    "internetservice": "DSL",
53    "techsupport": "No",
54    "contract": "1",
55    "churn": "Yes",
56    "seniorcitizen": "0",
57    "deviceprotection": "No",
58    "streamingtv": "No",
59    "totalcharges": "108.15",
60    "dependents": "0",
61    "monthlycharges": "53.85",
62    "customerid": "3668-OPV8K",
63    "dependents": "0",
64    "onlinebackup": "Yes",
65    "phoneservice": "Yes",
66    "tenure": "2",
67    "paymentmethod": "Mailed check"
68  }
]

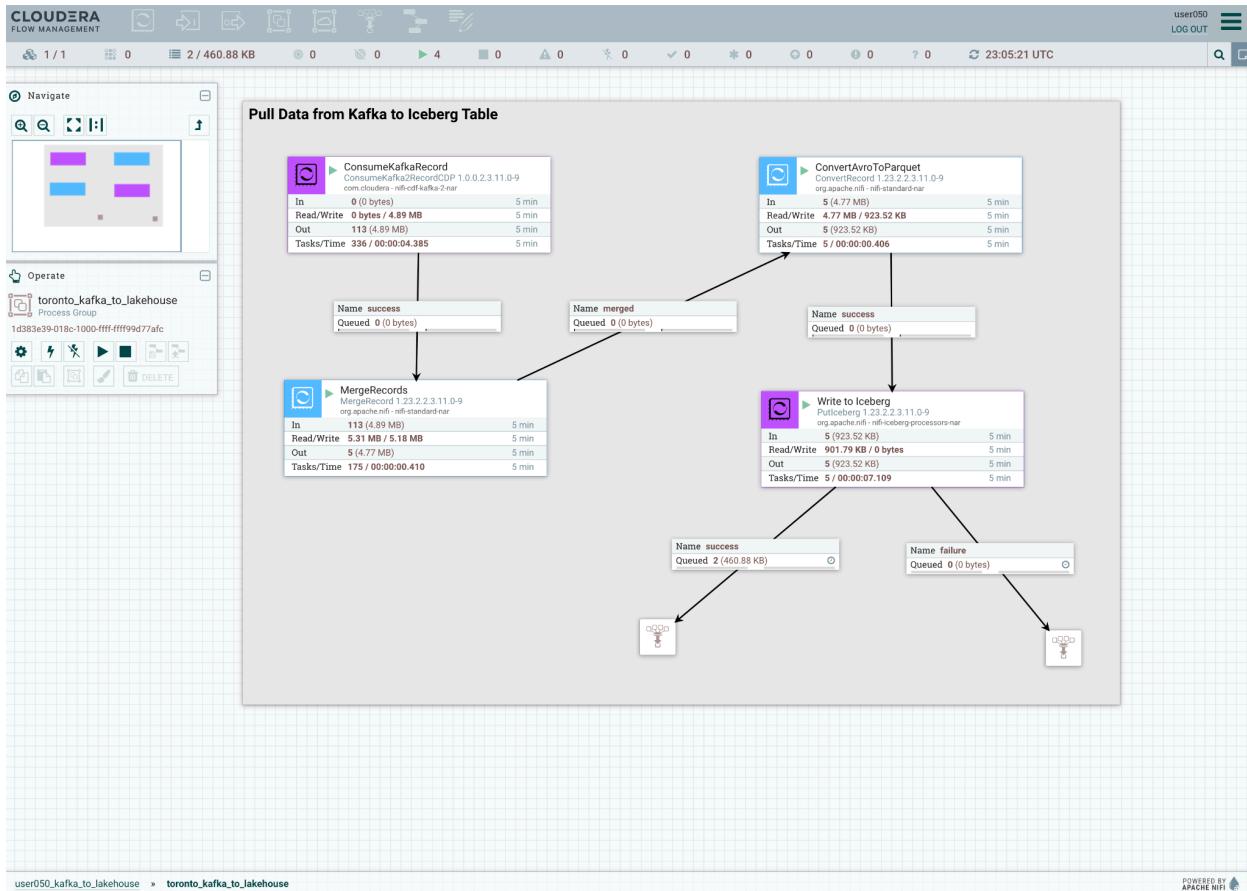
```

Filename: 2055d337-695f-4c6d-8203-3ece27a62dee  
Content Type: application/avro-binary

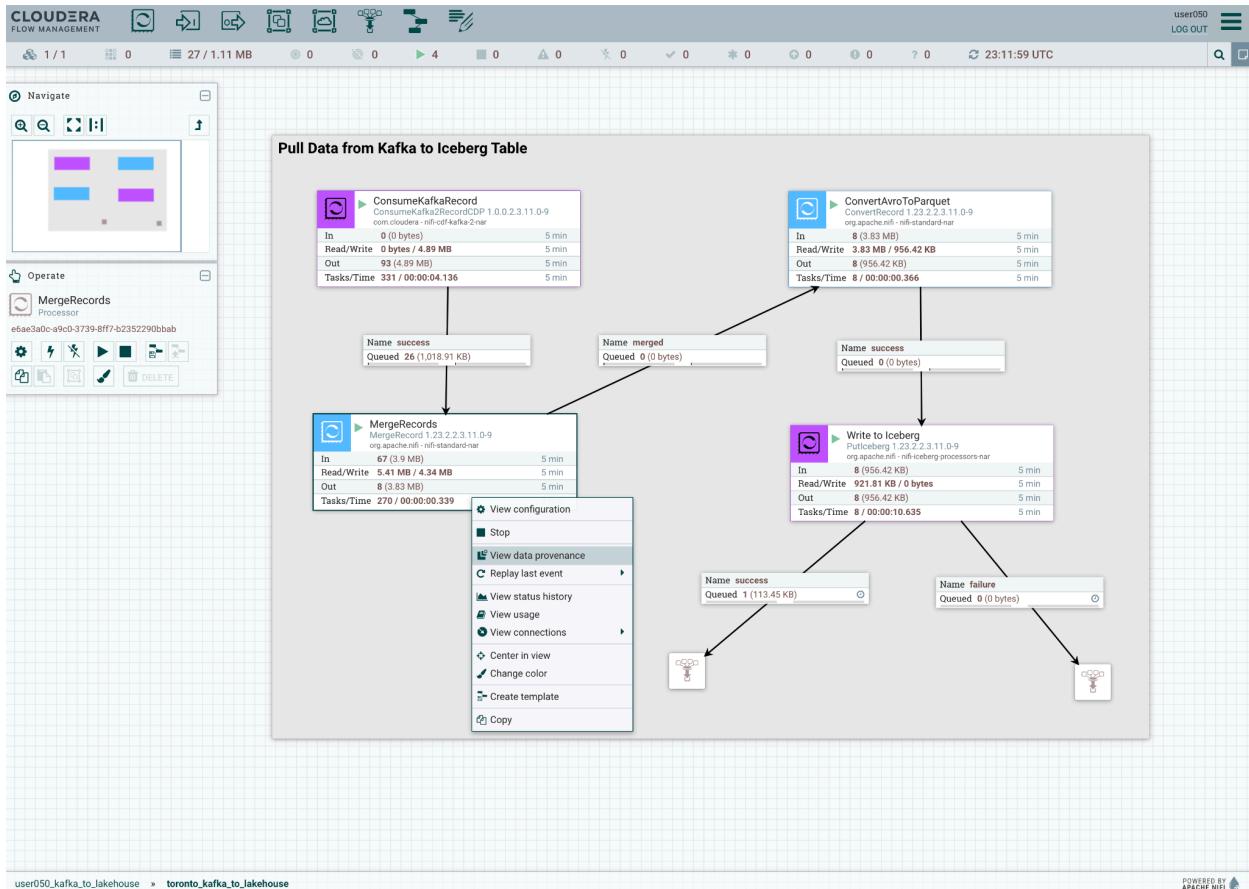
26. Start the stopped: **MergeRecords** processor again to resume the flow. Remember that you can refresh the flow counters with the combination Control+R or Command+R.



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



27. BONUS: NiFi is a powerful ingestion tool that gives you granular visibility into everything that's done to the data - for example, right-click on any processor and then click on **View data provenance** to see this in action



NiFi Data Provenance

Showing 989 of 989  
Oldest event available: 11/29/2023 22:46:09 UTC

Filter by component name

Date/Time ▾ Type FlowfileUuid Size Component Name Component Type Node

Date/Time	Type	FlowfileUuid	Size	Component Name	Component Type	Node
11/29/2023 23:14:46.125 UTC	DROP	42733ef5-db16-49b0-a4c5-b279146...	13.56 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	ee097a74-77b5-4a68-8398-22be85...	18.49 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	315a24eb-945d-41cf-b434-9e86075...	5.81 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	a3bca232-4acd-4393-98af-30aed5a...	7.75 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	92938e45-8a47-40ab-bfcf-b1a70748...	141 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	38c9253b-0e5b-463f-94db-2e50d86...	3.64 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	272009b5-e401-47da-9aa6-e00834...	88.16 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	e7555479-32ce-4cf8-9726-9105f615...	1.24 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	2e9254e9-a726-4456-925e-eae057...	97.71 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	df5ae4de-e123-4b0c-86b2-cea60b4...	20.22 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	29122cd-6744-491f-f447-6272e8...	24.77 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	DROP	de346e95-58ed-4428-9a27-1952eb...	41.84 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	42733ef5-db16-49b0-a4c5-b279146...	13.56 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	ee097a74-77b5-4a68-8398-22be85...	18.49 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	315a24eb-945d-41cf-b434-9e86075...	5.81 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	a3bca232-4acd-4393-98af-30aed5a...	7.75 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	92938e45-8a47-40ab-bfcf-b1a70748...	141 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	38c9253b-0e5b-463f-94db-2e50d86...	3.64 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	272009b5-e401-47da-9aa6-e00834...	88.16 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	e7555479-32ce-4cf8-9726-9105f615...	1.24 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	2e9254e9-a726-4456-925e-eae057...	97.71 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	df5ae4de-e123-4b0c-86b2-cea60b4...	20.22 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	29122cd-6744-491f-f447-6272e8...	24.77 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	ATTRIBUTES_MODIFIED	de346e95-58ed-4428-9a27-1952eb...	41.84 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:46.125 UTC	JOIN	4c206bbe-ba1f-42d0-9ba1-9a5d79...	451.87 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	DROP	468f3629-505a-4c06-914e-603a31b...	74.67 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	DROP	6799e6b8-84ce-40b6-baeb-19e1bcb...	78.54 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	DROP	7d818052-54a5-4b2e-a909-9fd7879...	115.51 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	DROP	a61953b0-87e4-4612-a642-d77c2...	124.73 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	DROP	7c70cd79-339a-4e98-a8d0-87077d...	81.42 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	DROP	c5701ca1-2197-485c-a6ea-5f22e24...	75.13 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	ATTRIBUTES_MODIFIED	468f3629-505a-4c06-914e-603a31b...	74.67 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	ATTRIBUTES_MODIFIED	6799e6b8-84ce-40b6-baeb-19e1bcb...	78.54 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	ATTRIBUTES_MODIFIED	7d818052-54a5-4b2e-a909-9fd7879...	115.51 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	ATTRIBUTES_MODIFIED	a61953b0-87e4-4612-a642-d77c2...	124.73 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...
11/29/2023 23:14:09.03 UTC	ATTRIBUTES_MODIFIED	7c70cd79-339a-4e98-a8d0-87077d...	81.42 KB	MergeRecords	MergeRecord	dfx-nifi-0.dfx-nifi.dfx-user050-kafka...

Last updated: 23:14:56 UTC

user050\_kafka\_to\_lakehouse » toronto\_kafka\_to\_lakehouse

APACHE NIFI