

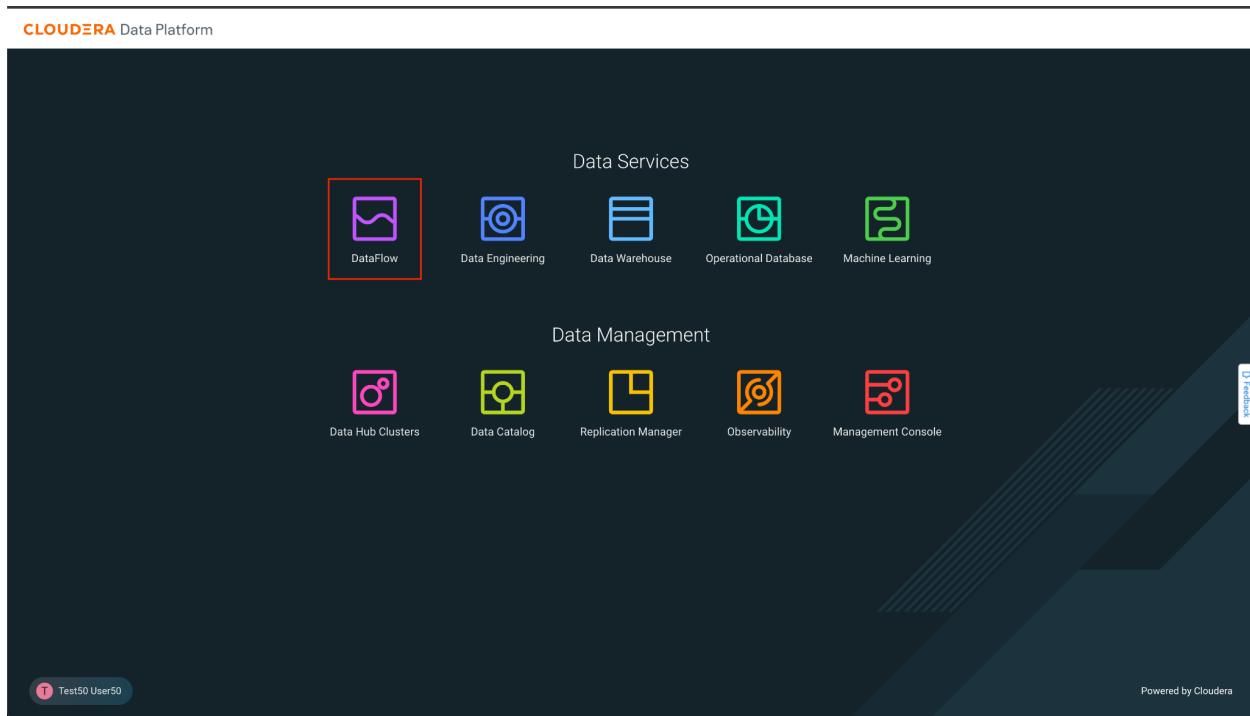
Data Lifecycle CDP Public Cloud

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 **toronto_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, Projects, Functions, and Environments. The main area is titled "Flow Catalog" and contains a search bar with the query "toronto_kafka". A table lists one flow entry:

Name	Type	Versions	Last Updated
toronto_kafka_to_lakehouse	Custom Flow Definition	3	3 minutes ago

At the bottom right of the catalog area, there are pagination controls: "Items per page: 10", "1 – 1 of 1", and navigation arrows. There is also a "Import Flow Definition" button and a "REFRESHED: 16 seconds ago" message.

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 (which is selected), ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and Test50 User50. The main area is titled 'Flow Catalog' and contains a search bar. A list of flows is shown, with 'kafka_to_lakehouse' highlighted. To the right of the flow list is a detailed view of the 'kafka_to_lakehouse' flow. It includes sections for 'Name' (kafka_to_lakehouse), 'Flow Description' (No description specified), 'CRN #' (crn:cdp:df:us-west-1:508fd88f-8076-498a-acfb-6f8765cd35e8:flow:kafka_to_lakehouse), and deployment details. A 'Deploy' button is visible at the bottom.

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 **tor-hol-cdp-env**. The workshop instructor will tell you which environment to select. Once selected, click **Continue**.

The screenshot shows a 'New Deployment' dialog box overlaid on the Cloudera DataFlow interface. The dialog has a title 'New Deployment' and a sub-instruction 'Select the target environment'. It includes a note about sensitive data and a 'Selected Flow Definition' section showing 'NAME: kafka_to_lakehouse' and 'VERSION: 1'. Below this is a 'Target Environment' dropdown menu where 'aws ssa-hol' is selected. At the bottom of the dialog are 'Cancel' and 'Continue →' buttons. The background shows the same flow catalog and deployment details as the previous screenshot.

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: user000_kafka_to_lakehouse
 Deployment name is valid

Selected Flow Definition

NAME	VERSION
toronto_kafka_to_lakehouse	3

Target Environment

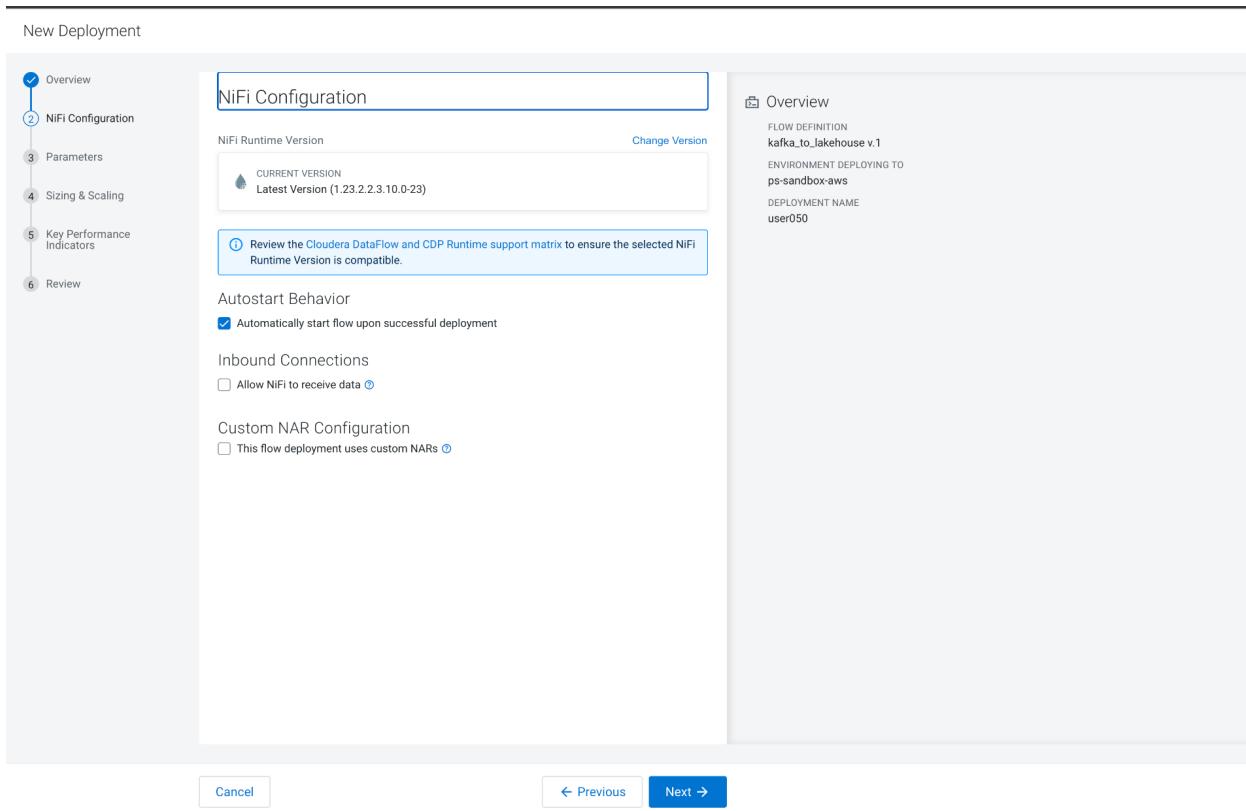
aws	NAME
aws	tor-hol-cdp-env

Target Project ⓘ

Select a project

Cancel Next →

6. Make sure the option **Automatically start flow upon successful deployment** is checked and click **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 number, *user050*, for example.

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
③ Parameters
4 Sizing & Scaling
5 Key Performance Indicators
6 Review

toronto_kafka_to_lakehouse (6)

CDPEnvironment ⓘ

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

Drop file or browse Select File

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.

kafka_brokers

streams-corebroker0.toronto.z30z-14kp.cloudera.site:9093,streams-corebroker1.toronto.z30z-14kp.cloudera.site:9093,streams-corebroker2.toronto.z30z-14kp.cloudera.site:9093

170/100K

kafka_topic

telco_data

10/100K

table

telco_iceberg_kafka

19/100K

workload_password

Enter parameter values.

0/100K

workload_user ⓘ

Enter parameter values.

0/100K

✓ Overview
FLOW DEFINITION toronto_kafka_to_lakehouse v.6
ENVIRONMENT DEPLOYING TO toronto-cdp-env
PROJECT ASSIGNING TO Workshop
DEPLOYMENT NAME user50_kafka_to_lakehouse

NiFi Configuration
NIFI RUNTIME VERSION Latest Version (1.23.2.2.3.11.0-9)
AUTO-START FLOW Yes
INBOUND CONNECTIONS No
CUSTOM NAR CONFIGURATION No

Cancel ← Previous Next →

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:

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](#)

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

Review

[View CLI Command](#)

- ✓ Overview
- ✓ NiFi Configuration
- ✓ Parameters
- ✓ Sizing & Scaling
- ✓ Key Performance Indicators
- 6 Review

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

[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*.

CLOUDERA DataFlow

Dashboard

Filter By: STATUS All - 15 ENVIRONMENTS All - 1

Status	Name ↑
Deploying	user050 ssa-hol

user050
ssa-hol

Deployment Initiated
Initiated deployment of [user050].

KPIs System Metrics Alerts

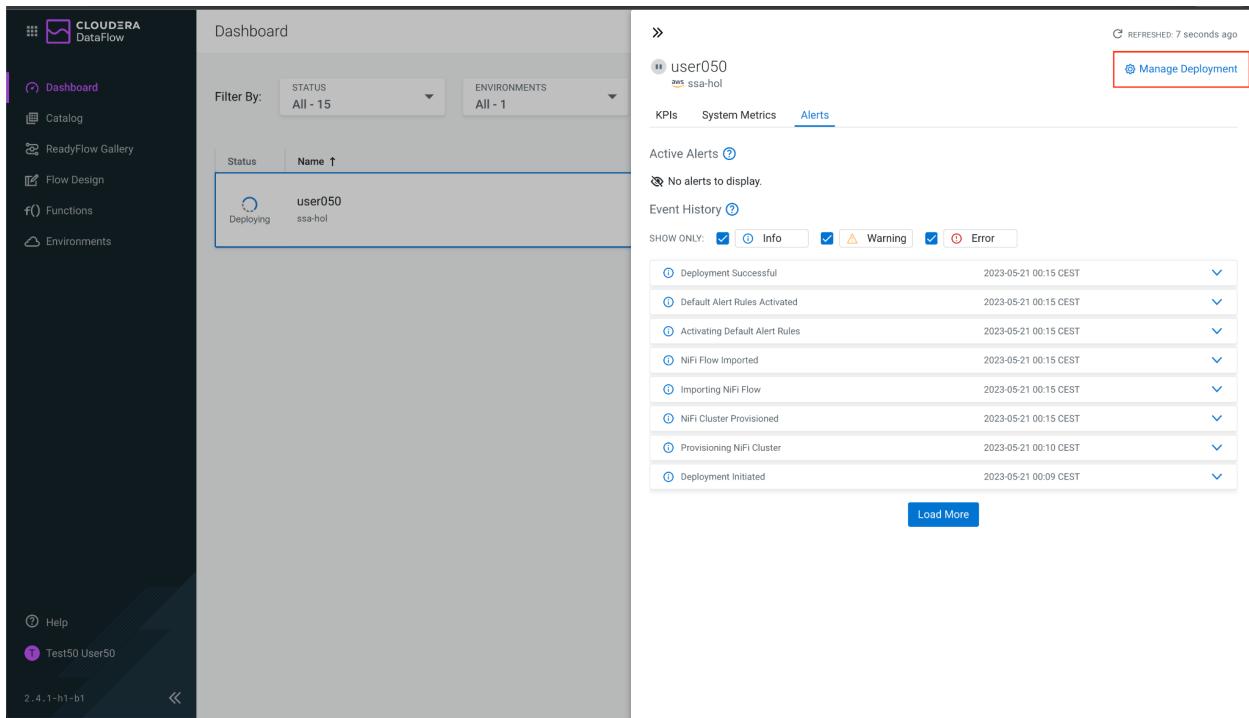
Active Alerts ⓘ
No alerts to display.

Event History ⓘ
SHOW ONLY: Info Warning Error

Deployment Initiated 2023-05-21 00:09 CEST

[Load More](#)

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 '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.

The screenshot shows the Cloudera DataFlow Deployment Manager interface. On the left, there's a sidebar with navigation links: Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, and Environments. The main area displays a deployment named 'user050' under 'Deployment Manager'. The deployment details include:

Category	Value
STATUS	Suspended
NODE COUNT	1
ENVIRONMENT	aws ssa-hol
DEPLOYMENT NAME	user050
AUTO SCALING	Disabled
REGION	US East(N. Virginia)
FLOW DEFINITION	kafka_to_lakehouse V.1
CREATED ON	2023-05-21 00:09 CEST
NIFI RUNTIME VERSION	1.20.0.2.3.8.2-2
DEPLOYED BY	Test50 User50
LAST UPDATED	2023-05-21 00:15 CEST
CRN #	CRM.CDP.DF.US.WES

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 deployment details, there's a button to 'Recreate Deployment CLI Command'. Under 'Deployment Settings', there are tabs for KPIs and Alerts (which is selected), Sizing and Scaling, Parameters, and NiFi Configuration. The 'Key Performance Indicators' section allows users to add new KPIs. At the bottom, there are buttons for Discard Changes, Apply Changes, and Update Deployment CLI Command.

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

The screenshot shows the Cloudera Flow Management interface. At the top, there's a toolbar with various icons for operations like Create, Delete, Copy, Paste, and Filter. The top right corner shows the user 'acampos' and a 'LOG OUT' button. The top status bar displays metrics: 1 / 1, 0 / 0 bytes, 0 / 0 bytes, 4 / 0, 0 / 0, 0 / 0, 0 / 0, 0 / 0, 0 / 0, 0 / 0, 0 / 0, and the timestamp 10:33:37 UTC.

On the left, there's a sidebar titled 'Navigate' with search and filter tools. Below it is a large empty workspace area.

The main content area features a 'processGroup' window. This window has a header with metrics: 0 / 0 bytes, 0 / 0 bytes, 4 / 0, 0 / 0, and 0 / 0. It contains a table with four rows:

	Queued	In	Read/Write	Out
0	0 (0 bytes)	0 (0 bytes) → 0	0 bytes / 0 bytes	0 → 0 (0 bytes)
0				
0				

Below the table is a footer with metrics: 0 / 0 bytes, 0 / 0 bytes, 0 / 0, 0 / 0, and 0 / 0.

On the far left, under the 'Operate' section, there's a node named 'user050' which is identified as a 'Process Group'. Its ID is 476b1aca-011b-1000-ad41-c3a80b93f1b6. Below this are several small icons for node operations: Create, Delete, Copy, Paste, and others.

At the bottom left, there's a link to another 'user050' node. The bottom right corner has a 'POWERED BY APACHE NIFI' watermark.

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:

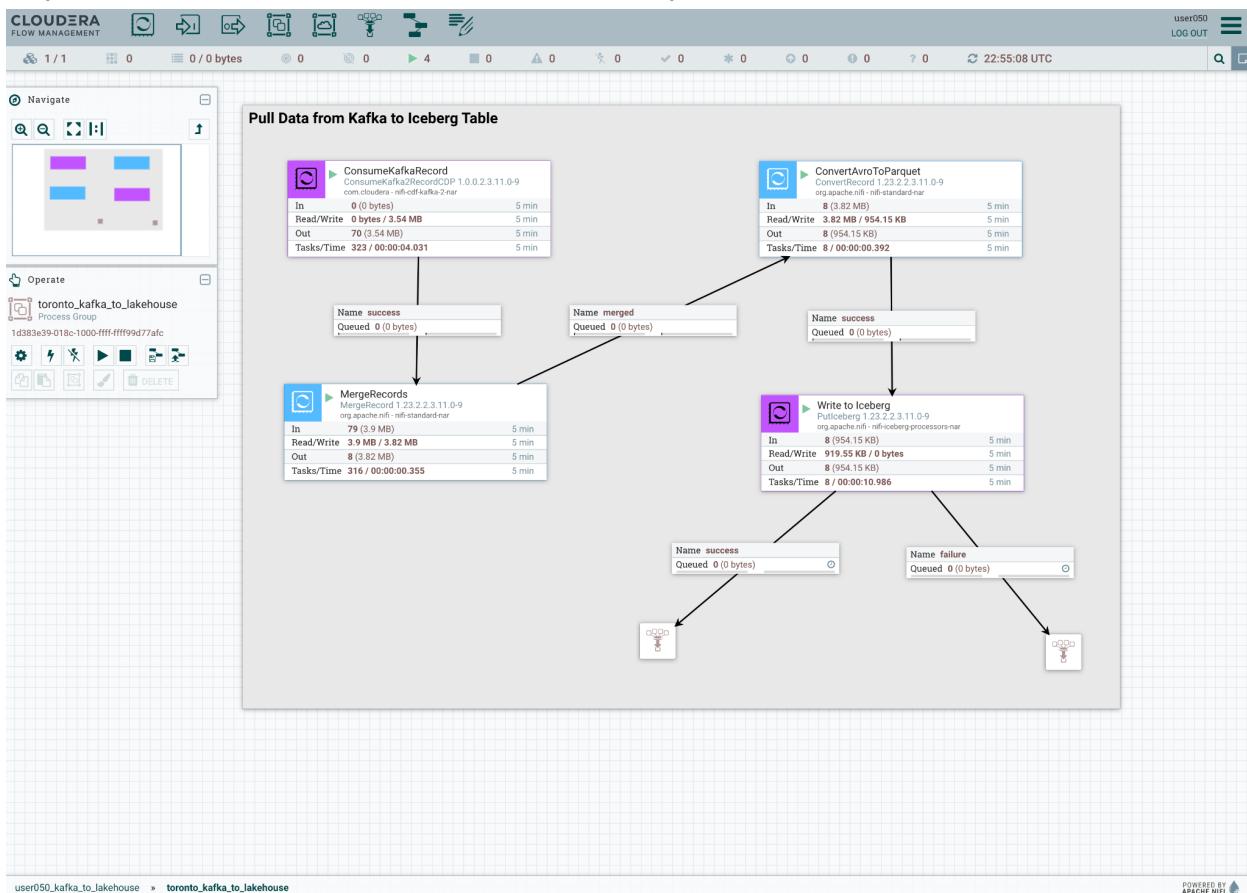
ConsumeKafkaRecord, consumes data from the Kafka topic, reading the data in JSON and outputting in AVRO.

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).

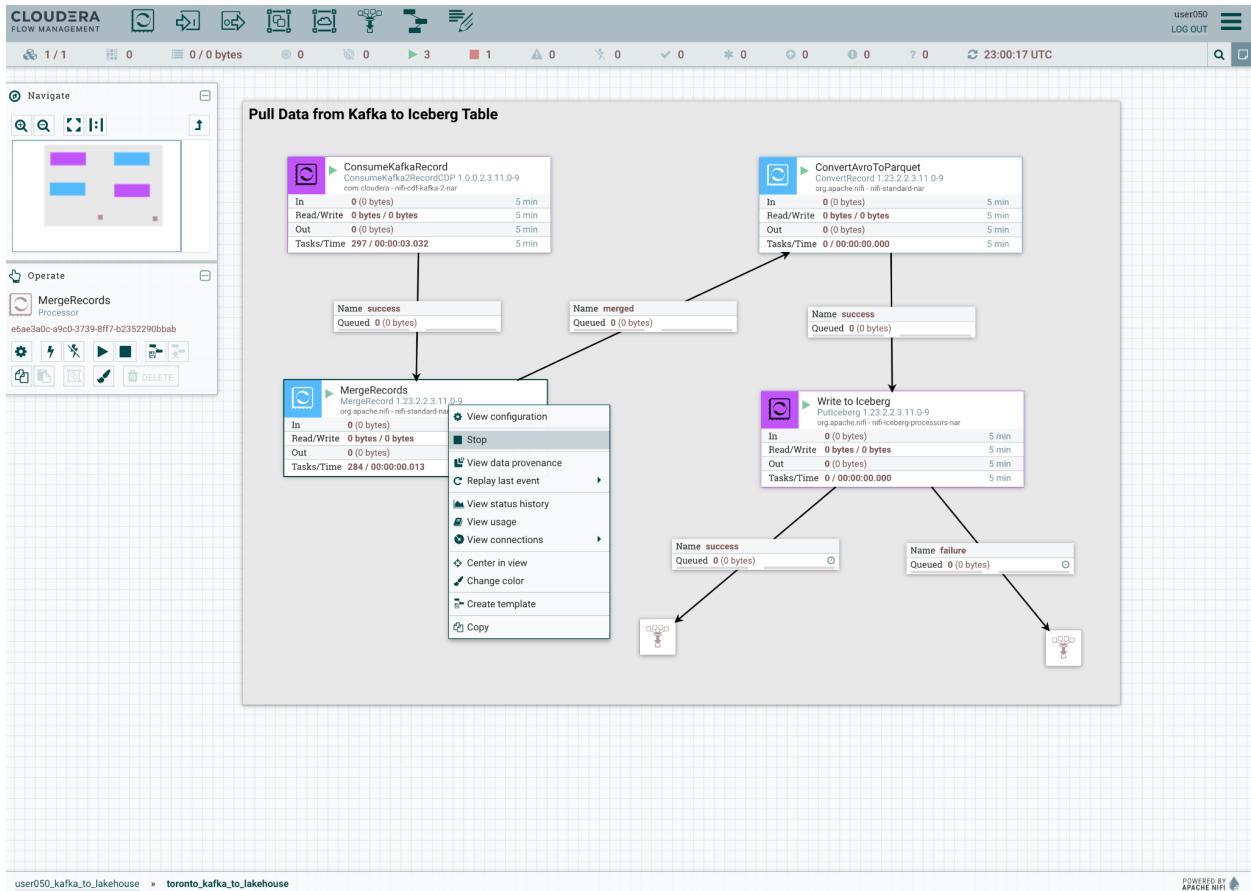
As you can see, the Processors are not started, they are paused.



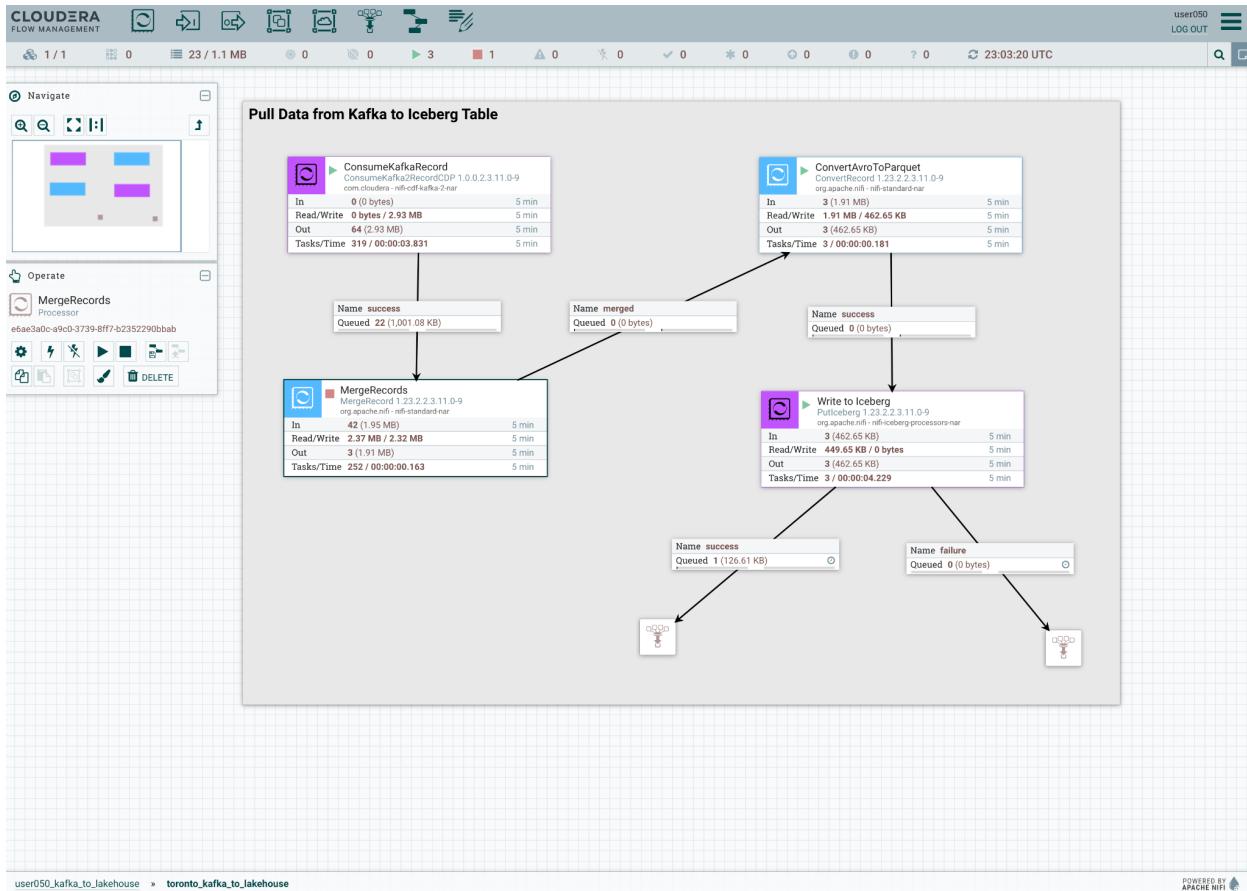
From the Out field in every processor, you can see that data has flowed through in the past 5 minutes. You have already consumed data from Kafka and to Iceberg!

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. First, right-click on **MergeRecords** processor and click **Stop**.

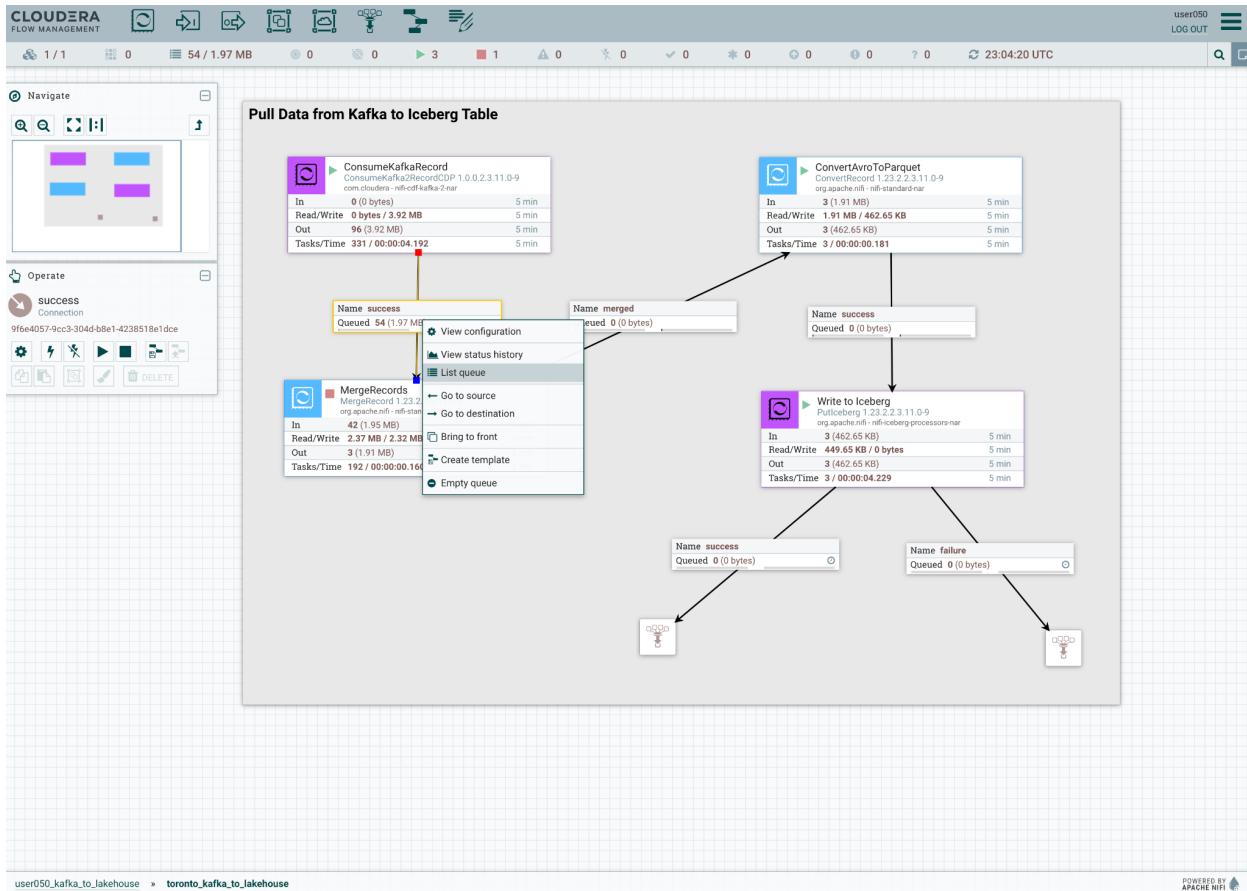


18. You will see data start to queue up in the connector shortly after stopping the MergeRecords processor.



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.



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".

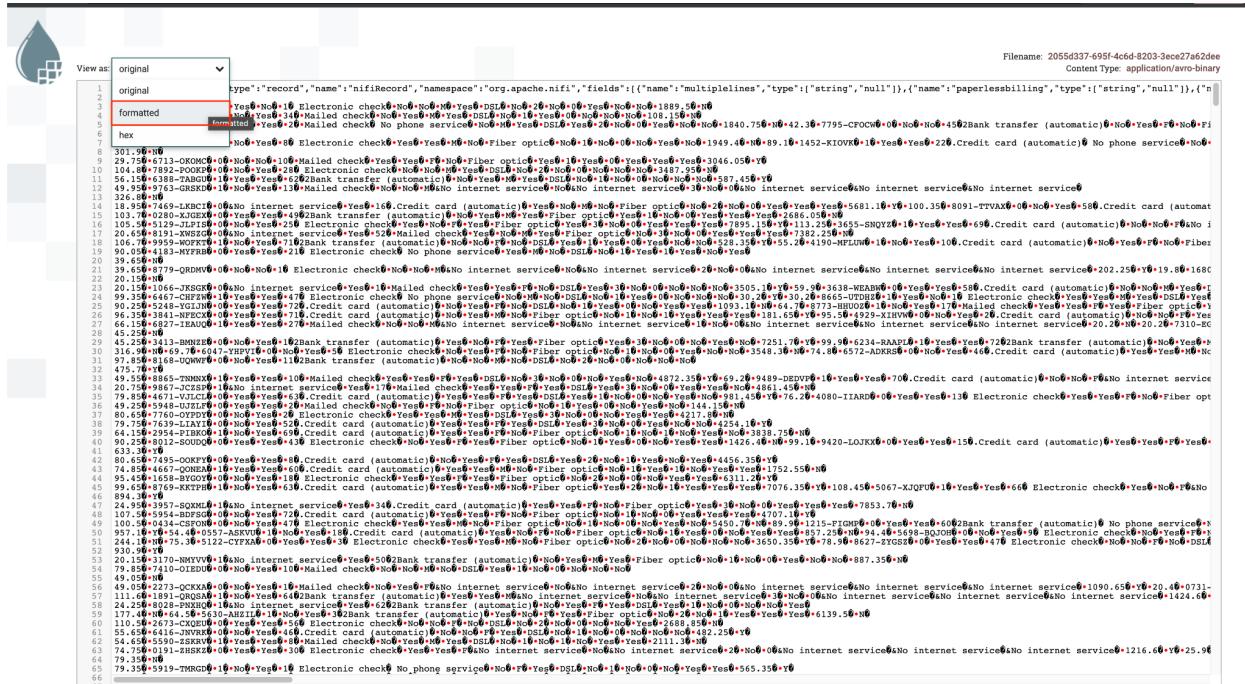
20. Once the FlowFile detail window appears, click on the button **VIEW** to open the content of consumed events.

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	DOWNLOAD VIEW
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**.



	original	formatted
1	original	type: "record", "name": "nifiRecord", "namespace": "org.apache.nifi", "fields": [{"name": "multiplelines", "type": ["string", "null"]}, {"name": "paperlessbilling", "type": ["string", "null"]}], "n
2		o: No* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* No* 18
3		20 Mailed check* No* No* Yes* 19 Bank transfer (automatic)* No* Yes* 18
4		21 Mailed check* No* No* Yes* 19 Bank transfer (automatic)* No* Yes* 18
5		22 Mailed check* No* No* Yes* 19 Bank transfer (automatic)* No* Yes* 18
6		23 Mailed check* No* No* Yes* 19 Bank transfer (automatic)* No* Yes* 18
7		24 Mailed check* No* No* Yes* 19 Bank transfer (automatic)* No* Yes* 18
8	301.50-100	No* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
9	301.50-11-OKRNG	No* Yes* 19 Mailed check* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
10	104.80-7892-POOKN	No* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
11	56.15-6388-TABGU	19 Yes* Yes* 19 Credit card (automatic)* No* Yes* 18
12	104.80-1063-GREKX	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
13	326.80-#	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
14	18.50-7469-LKRC19	No* Yes* 19 Credit card (automatic)* No* Yes* 18
15	18.50-7469-LKRC19	No* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
16	105.50-5129-JLP10	19 No* Yes* 19 Electronic check* No* Yes* 18 Credit card (automatic)* No* Yes* 18
17	20.65-8191-XNSKD9	No* Yes* 19 Credit card (automatic)* No* Yes* 18
18	105.50-5129-JLP10	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
19	90.05-4183-NYFRB	19 Yes* Yes* 19 Credit card (automatic)* No* Yes* 18
20	39.65-#	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
21	105.50-8779-QRDWQ	19 No* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
22	20.15-#	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
23	105.50-6666-JRSGR	No* Yes* 19 Mailed check* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
24	99.35-6467-CFZFD	19 Yes* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
25	90.25-5248-YGLJN	19 Yes* Yes* 19 Credit card (automatic)* No* Yes* 18
26	105.50-5129-JLP10	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
27	65.15-6827-IAUDQ	19 Yes* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
28	45.25-#	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
29	105.50-13-BHNU18	19 No* Yes* 19 Bank transfer (automatic)* No* Yes* 18 Credit card (automatic)* No* Yes* 18
30	316.90-#69.78-#047-HPFT	19 No* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
31	97.85-8168-UQHJN	19 No* Yes* 11#2 Bank transfer (automatic)* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
32	49.55-#	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
33	49.55-8865-TNNX9	19 Yes* Yes* 19 Mailed check* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
34	25.75-9867-JCZSF	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
35	105.50-5129-JLP10	No* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
36	49.25-5948-UJZLDR	19 Yes* Yes* 19 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
37	89.65-#0777-OYVZD	19 Yes* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
38	105.50-7639-PAVAY	19 Yes* Yes* 19 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
39	64.15-2954-P1BKQ	19 No* Yes* 18 Credit card (automatic)* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
40	105.50-12-SOUDQG	19 Yes* Yes* 18 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
41	633.30-#	No* Yes* 18 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
42	80.65-7495-OKFYN	19 Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
43	105.50-5129-JLP10	No* Yes* 18 Credit card (automatic)* No* Yes* 18
44	95.45-1658-BVGOY	19 No* Yes* 18 Electronic check* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
45	99.65-8769-KXTPR	19 No* Yes* 18 Credit card (automatic)* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
46	49.55-#	No* Yes* 18 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
47	24.95-3957-SQXMD	19 No* Internet service* Yes* 14#2 Credit card (automatic)* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
48	107.50-5934-BDFSG	19 No* Yes* 18 Credit card (automatic)* No* Yes* 18 Credit card (automatic)* No* Yes* 18
49	105.50-5129-JLP10	No* Yes* 18 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
50	95.71-#0-54.48-0557-ASKTU	19 No* Yes* 18 Credit card (automatic)* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
51	244.00-#75.36-#122-CYTKA	19 Yes* Yes* 18 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
52	25.75-#	No* Yes* 18 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
53	20.15-3170-NMVV9	19 No* Internet service* 2#2 Bank transfer (automatic)* No* Yes* 18 Credit card (automatic)* No* Yes* 18
54	105.50-110-DIEODU	19 No* Yes* 18 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
55	49.05-#	No* Yes* 18 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
56	49.05-2273-QCKXKA	19 No* Yes* 18 Internet service* No* Internet service* 2#2 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
57	105.50-5129-JLP10	No* Yes* 18 Internet service* 2#2 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
58	24.25-8028-PNKHO	19 No* Internet service* Yes* 18#2 Bank transfer (automatic)* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
59	177.40-#0-64.58-#630-AHZLW	19 No* Yes* 18 Credit card (automatic)* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
60	105.50-5129-JLP10	No* Yes* 18 Mailed check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
61	55.65-6416-JNVRK	19 Yes* Yes* 18 Credit card (automatic)* No* Yes* 18 Credit card (automatic)* No* Yes* 18
62	54.65-5590-ZSRKV	19 Yes* Yes* 18 Mailed check* No* Yes* 18 Credit card (automatic)* No* Yes* 18
63	105.50-5129-ERHZE	No* Yes* 18 Electronic check* Yes* Yes* 18 Credit card (automatic)* No* Yes* 18
64	79.35-#	No* Yes* 18 Internet service* 2#2 No* Internet service* 2#2 No* Internet service* 2#2 No* Internet service* 2#2
65	79.35-5919-TMRGDW	19 No* Yes* 18 Electronic check* No* No* Yes* 18 Credit card (automatic)* No* Yes* 18
66	79.35-#	No* Yes* 18 Internet service* 2#2 No* Internet service* 2#2 No* Internet service* 2#2 No* Internet service* 2#2
67	79.35-#	No* Yes* 18 Internet service* 2#2 No* Internet service* 2#2 No* Internet service* 2#2 No* Internet service* 2#2

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.

View as: formatted

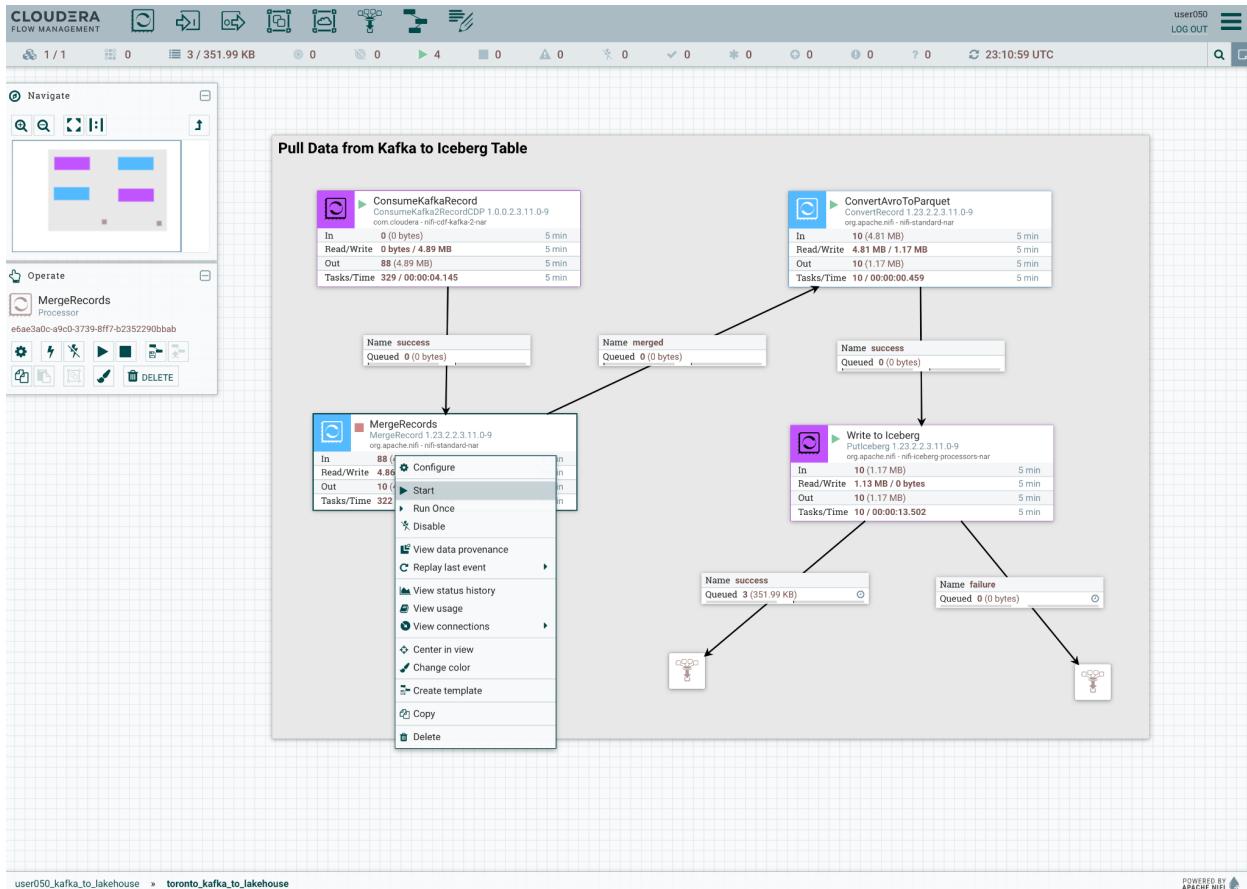
```

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-VV8EG",
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-QN9DE",
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    "partner": "No",
61    "customerid": "53.85",
62    "dependents": "0",
63    "monthlycharges": "53.85",
64    "customerid": "3668-OPV8K",
65    "phoneservice": "Yes",
66    "tenure": "2",
67    "paymentmethod": "Mailed check"
68  }
]

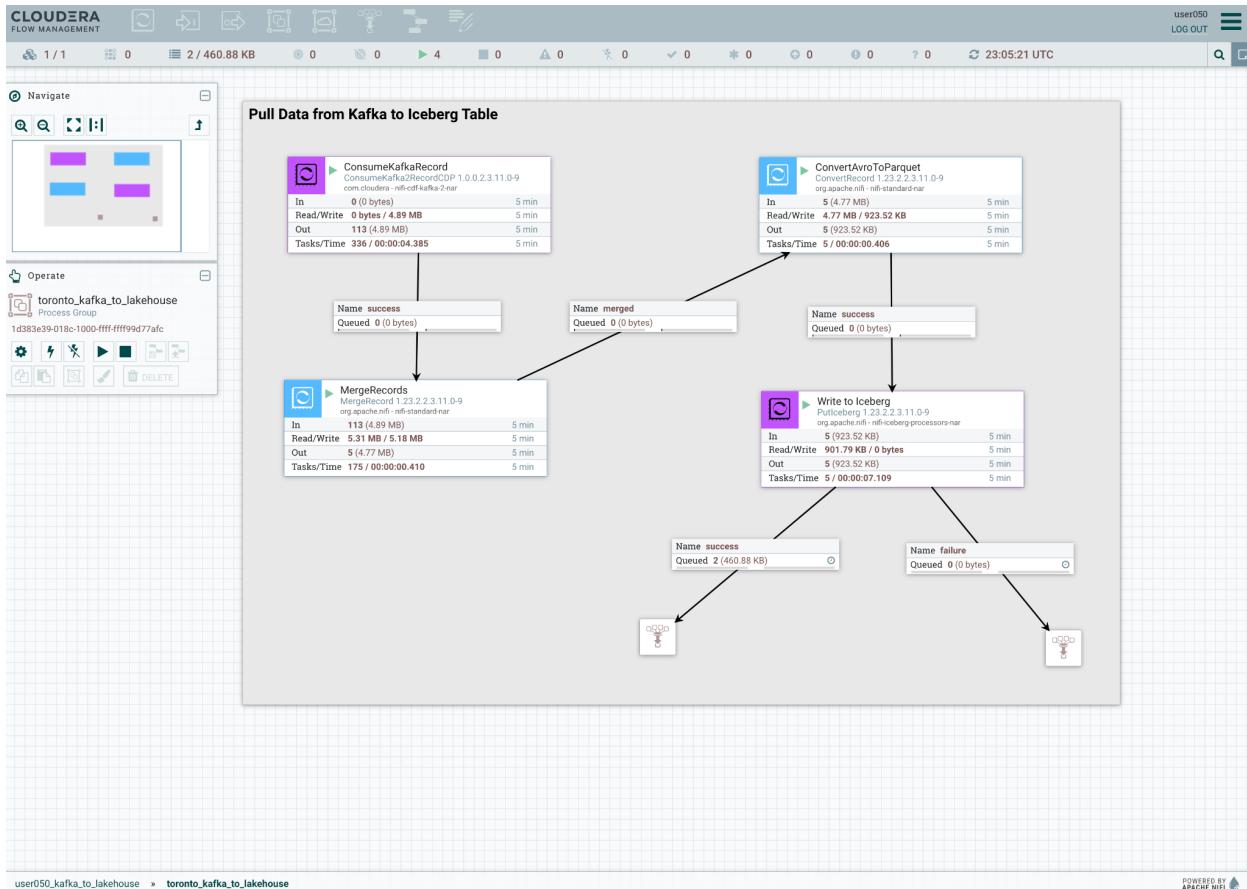
```

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

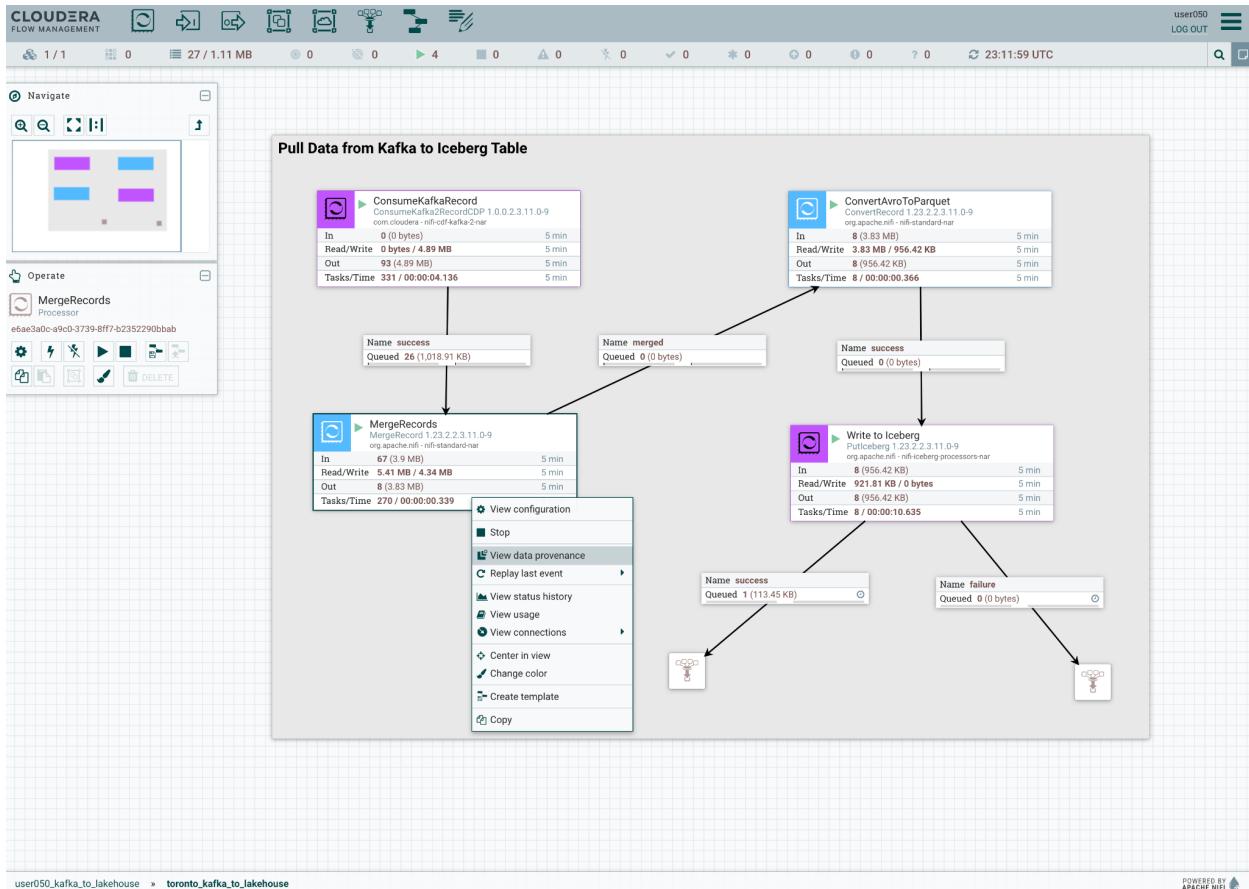
23. Start the stopped: **MergeRecords** processor again to resume the flow. Remember that you can refresh the flow counters with the combination Control+R or Command+R.



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



24. 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	ee0973a7477b5d4e8-8398-22be85...	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	a3bcfa232-4acd-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	92938ea5-84a7-40ab-bfcf-b1a0748...	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	e7555479-32ca-4cf8-9f26-9105f615...	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	ee0973a7477b5d4e8-8398-22be85...	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	a3bcfa232-4acd-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	92938ea5-84a7-40ab-bfcf-b1a0748...	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	e7555479-32ca-4cf8-9f26-9105f615...	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	7c70dd79-339a-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-a6e9-5f22e24...	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	7c70dd79-339a-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

EXPERIMENTAL APACHE NIFI