

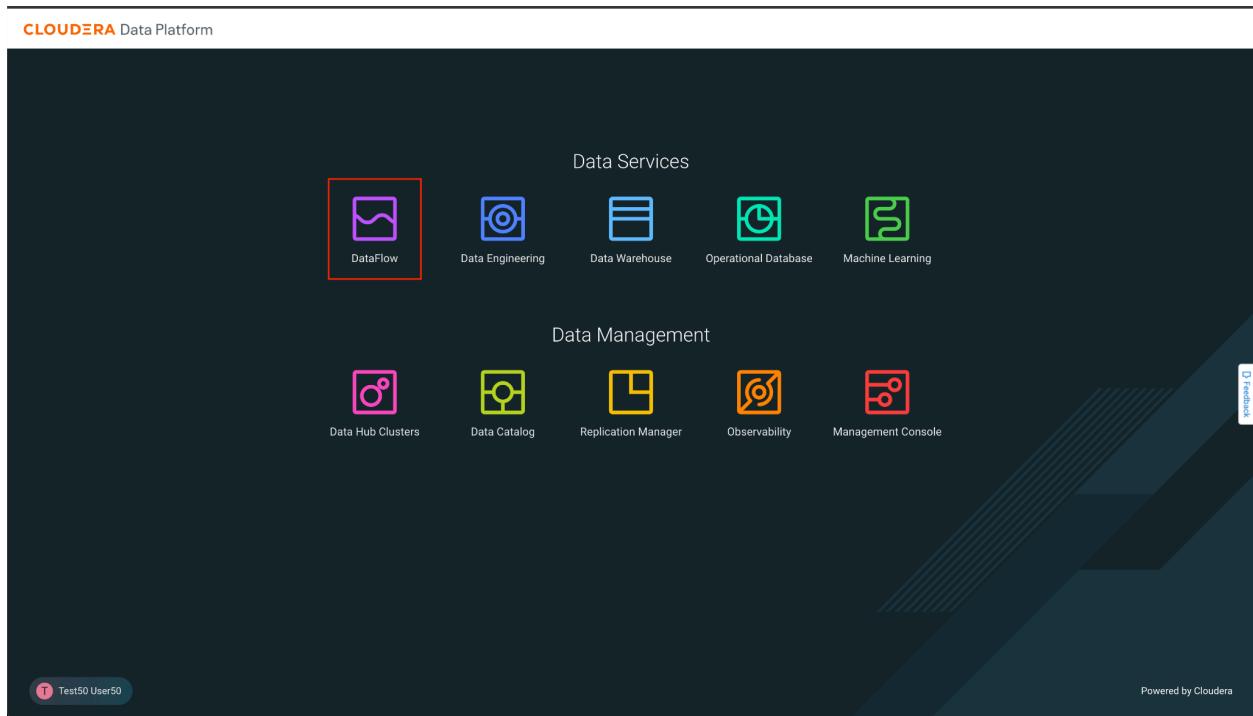
Data Lifecycle CDP Public Cloud

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.

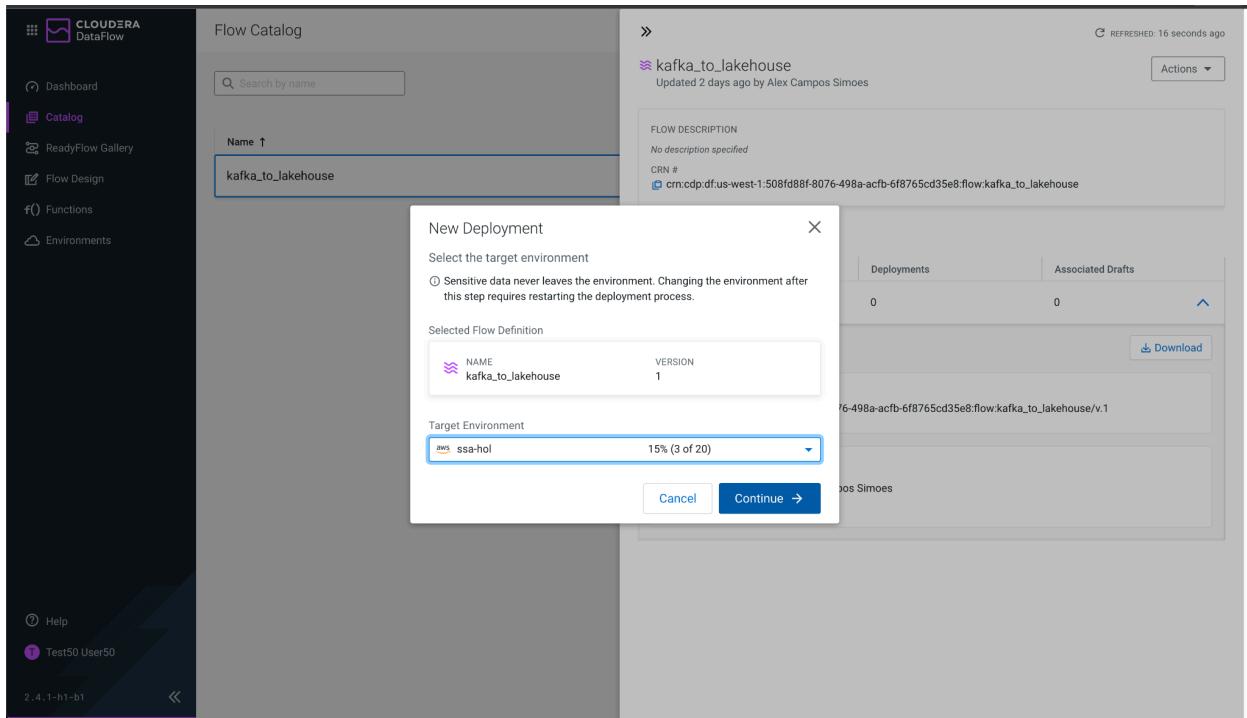
The screenshot shows the Cloudera DataFlow interface. On the left, a dark sidebar menu includes options like Dashboard, Catalog (which is selected and highlighted in purple), ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and a user profile for 'Test50 User50'. The main content area is titled 'Flow Catalog' and displays a table of flows. A search bar at the top of the table says 'Search by name'. The table has columns for Name, Type, Versions, and Last Updated. One row is visible: 'kafka_to_lakehouse' (Custom Flow Definition, 1 version, last updated 2 days ago). At the bottom right of the table, there are pagination controls ('Items per page: 10', '1 - 1 of 1', and arrows) and a refresh button ('REFRESHED: 5 seconds ago').

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, there's 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 shows a search bar with 'Search by name'. A list of flows is displayed, with 'kafka_to_lakehouse' selected and highlighted in blue. The right side provides detailed information about this flow:

- » kafka_to_lakehouse** (purple icon)
Updated 2 days ago by Alex Campos Simoes
REFRESHED: 7 seconds ago
- Actions** (dropdown menu)
- FLOW DESCRIPTION**
No description specified
- CRN #**
crm:cdp:df:us-west-1:508fd88f-8076-498a-acfb-6f8765cd35e8:flow:kafka_to_lakehouse
- Only show deployed versions
- | Version | Deployments | Associated Drafts |
|---------|-------------|-------------------|
| 1 | 0 | 0 |
- Deploy →** (blue button)
- Download** (button with download icon)
- CRN #**
crm:cdp:df:us-west-1:508fd88f-8076-498a-acfb-6f8765cd35e8:flow:kafka_to_lakehouse/v.1
- 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.

You can safely ignore the Warning

New Deployment

The screenshot shows the 'New Deployment' wizard in progress, specifically the 'Overview' step (step 1). The left sidebar lists steps 1 through 6: Overview, NiFi Configuration, Parameters, Sizing & Scaling, Key Performance Indicators, and Review. Step 1 is highlighted with a blue circle. The main area displays the following fields:

- Deployment Name:** user050 (highlighted with a blue border)
- Selected Flow Definition:** Kafka_to_lakehouse (NAME: Kafka_to_lakehouse, VERSION: 2)
- Target Environment:** env-israel-1 (aws icon)
- Target Project:** Unassigned (dropdown menu)
- Warning:** A yellow warning box states: "No Projects are associated with this workspace. Selecting 'Unassigned' will make this Deployment available to all DFFlowAdmin and DFFlowUser in env-israel-1."
- Import Configuration:** A note: "If you have previously exported a deployment configuration that closely aligns with this one, you can import it here to auto-fill as much of the wizard as possible." with an 'Import' button.

At the bottom are 'Cancel' and 'Next →' buttons.

New Deployment

1 Overview

2 NI FI Configuration

3 Parameters

4 Sizing & Scaling

5 Key Performance Indicators

6 Review

Overview

Deployment Name: user050 (Deployment name is valid)

Selected Flow Definition: NAME kafka_to_lakehouse VERSION 1

Target Environment: NAME ssa-hol

[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

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 [?](#)

Cancel

← Previous

Next →

Overview

FLOW DEFINITION
kafka_to_lakehouse v.1

ENVIRONMENT DEPLOYING TO
ps-sandbox-aws

DEPLOYMENT NAME
user050

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

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 17/100K

CDP Workload Username 7/100K

user050

CDPEnvironment 0/100K

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

Select File Drop file or browse

① DataFlow automatically adds all required configuration files to interact with Data Lake services. Unnecessary files that are added won't impact the deployment process.

Cancel Previous Next →

See the next page for the rest of the configurations

New Deployment

1 Overview

2 NiFi Configuration

3 Parameters

4 Sizing & Scaling

5 Key Performance Indicators

6 Review

CDPEnvironment

core-site.xml (green)
ssl-client.xml (green)
hive-site.xml (green)

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.

Database

user050

7/100K

Kafka Brokers

realtime-ingestion-corebroker0.ssa-hol.yu11-vbzg.cloudera.site:9093,realtime-ingestion-corebroker1.ssa-hol.yu11-vbzg.cloudera.site:9093,realtime-ingestion-corebroker2.ssa-hol.yu11-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**.

The screenshot shows the 'New Deployment' wizard at step 4: Sizing & Scaling. On the left, a sidebar lists steps 1 through 6: Overview, NiFi Configuration, Parameters, Sizing & Scaling (selected), Key Performance Indicators, and Review. The main panel is titled 'Sizing & Scaling' with the sub-section 'NiFi Node Sizing'. It displays four node size options: Extra Small, Small, Medium, and Large, each with its resource requirements. Below this is a section for 'Number of NiFi Nodes' with a slider set to 1. To the right, there's an 'Overview' section with flow details like 'kafka_to_lakehouse v.1' and environment 'ssa-hol', and a 'NiFi Configuration' section listing runtime settings. At the bottom are 'Cancel', 'Previous', and 'Next' buttons.

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

The screenshot shows the 'New Deployment' wizard at step 5: Key Performance Indicators. The sidebar is identical to the previous step. The main panel has a 'Key Performance Indicators' section with a note about setting up KPIs and a 'Learn more' link, followed by a dashed box containing an 'Add New KPI' button. To the right is the familiar 'Overview' and 'NiFi Configuration' sections. At the bottom are 'Cancel', 'Previous', and 'Next' buttons.

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

New Deployment

Review

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

CDP WORKLOAD USER PASSWORD
[Sensitive Value Provided]

CDP WORKLOAD USERNAME
user050

CDPENVIRONMENT

core-site.xml
ssi-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*.

The screenshot shows the Cloudera DataFlow interface. On the left, a sidebar menu includes options like Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and a user profile for Test50 User50. The main area is titled 'Dashboard' with filters for STATUS (All - 15) and ENVIRONMENTS (All - 1). A table lists flows by Status (Deploying) and Name (user050). To the right, a detailed view for flow 'user050' (aws ssa-hol) shows KPIs, System Metrics, and Alerts. An alert message is highlighted with a red box: 'Deployment Initiated' (Info level) - Initiated deployment of [user050]. Below this, the Event History shows a single entry: '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 dashboard. On the left, there's a 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 of flows. One flow is highlighted: "user050" (aws ssa-hol), which is currently "Deploying". To the right of the table, there's a section titled "user050" with tabs for KPIs, System Metrics, and Alerts (which is selected). Below this, it says "Active Alerts" and "No alerts to display". Under "Event History", there's a table of events:

Event	Timestamp
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 event history section.

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.

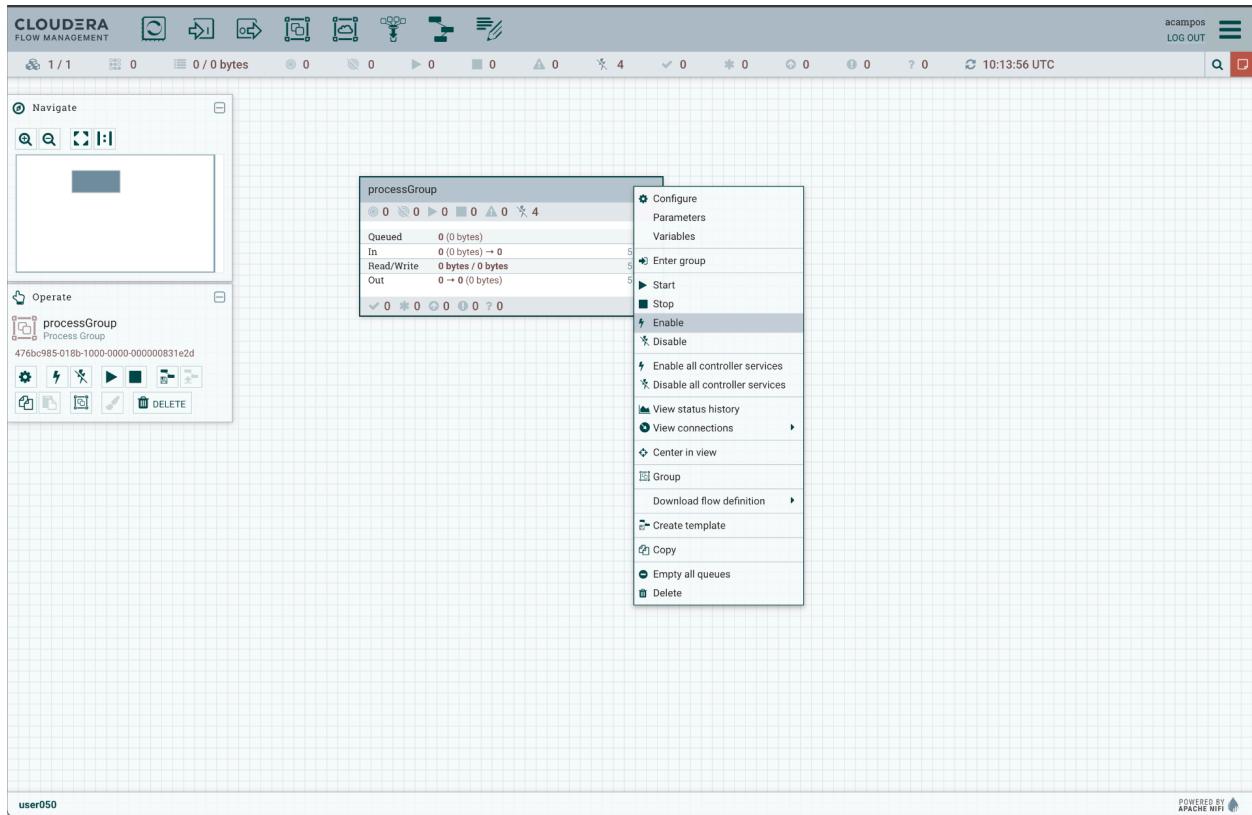
The screenshot shows the Cloudera DataFlow Deployment Manager interface. On the left, there's a sidebar with icons for Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and a user profile for Test50 User50. The main content area has a header "Deployment Manager" with a back link. Below it, a table displays deployment details:

STATUS	DEPLOYMENT NAME	FLOW DEFINITION	DEPLOYED BY
Suspended	user050	kafka_to_lakehouse V1	Test50 User50
NODE COUNT	AUTO SCALING	CREATED ON	LAST UPDATED
1	Disabled	2023-05-21 00:09 CEST	2023-05-21 00:15 CEST
ENVIRONMENT	REGION	NIFI RUNTIME VERSION	CRN #
aws ssa-hol	US East(N. Virginia)	1.2.0.2.3.8.2-2	crm:cdp:dfus-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 table is a button to "Recreate Deployment CLI Command".

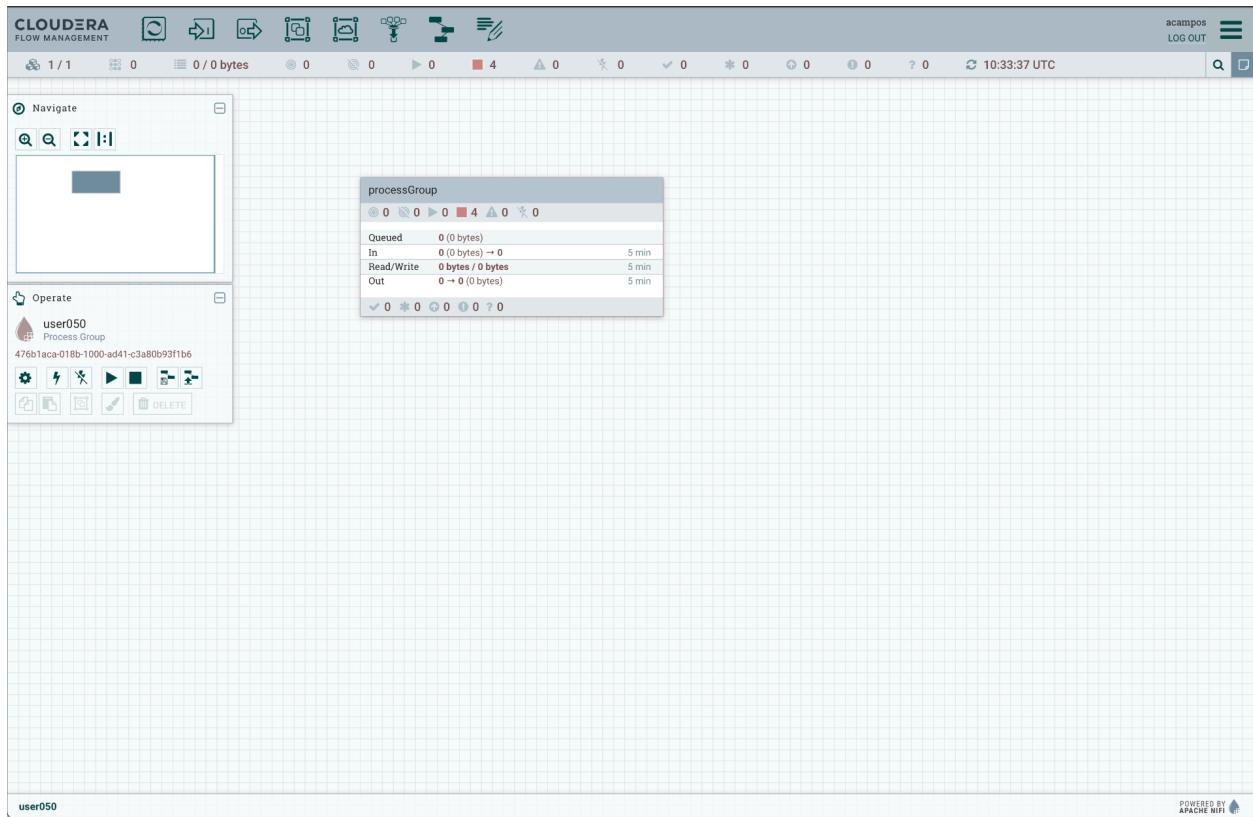
The "Deployment Settings" section includes tabs for KPIs and Alerts, Sizing and Scaling, Parameters, and NIFI Configuration. Under "Key Performance Indicators", there's a note to "Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed." A "Learn more" link is provided. At the bottom, there are "Discard Changes" and "Apply Changes" buttons, along with a "Update Deployment CLI Command" link.

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**.



Normally we would not need to enable the flow manually, it would already be running, this is just for learning and demonstration purposes.

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:

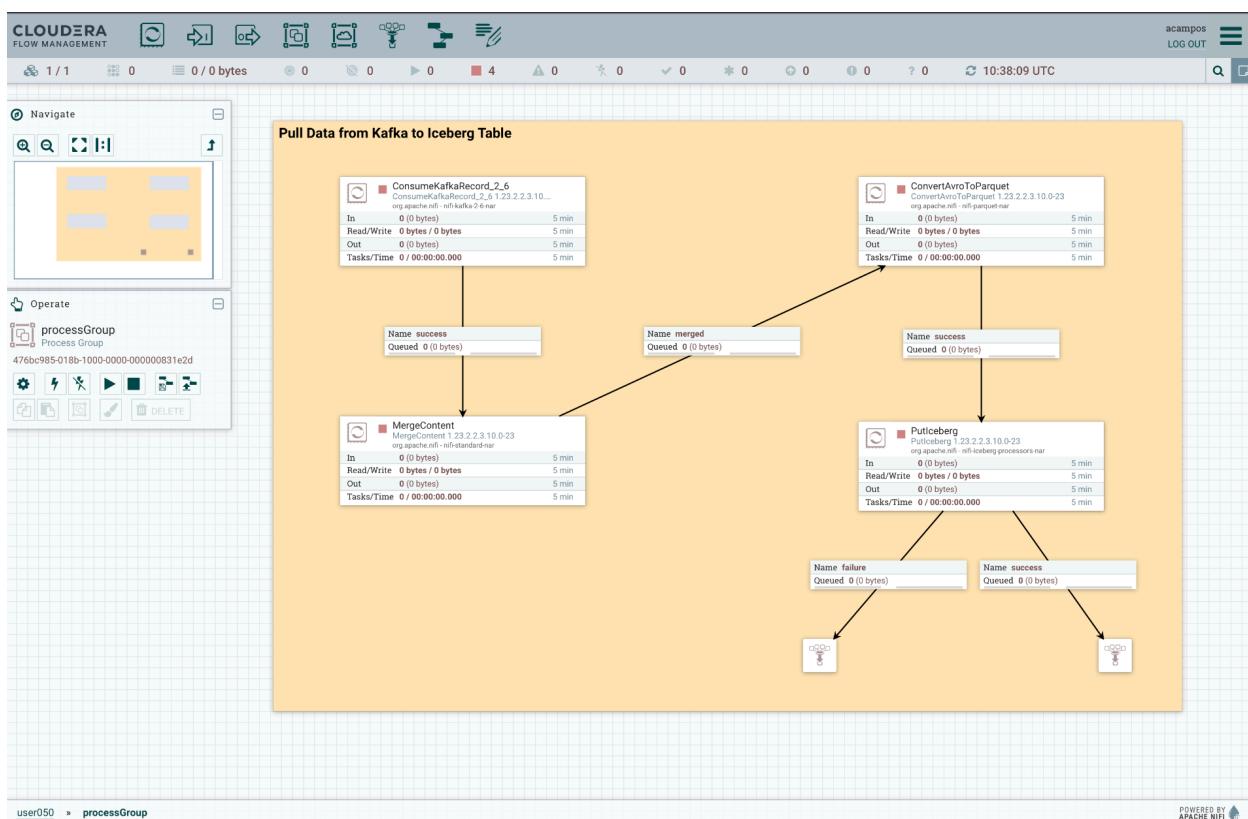
ConsumeKafkaRecord, processor to consume data from the Kafka topic, reading the data in JSON format and outputting in AVRO format.

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

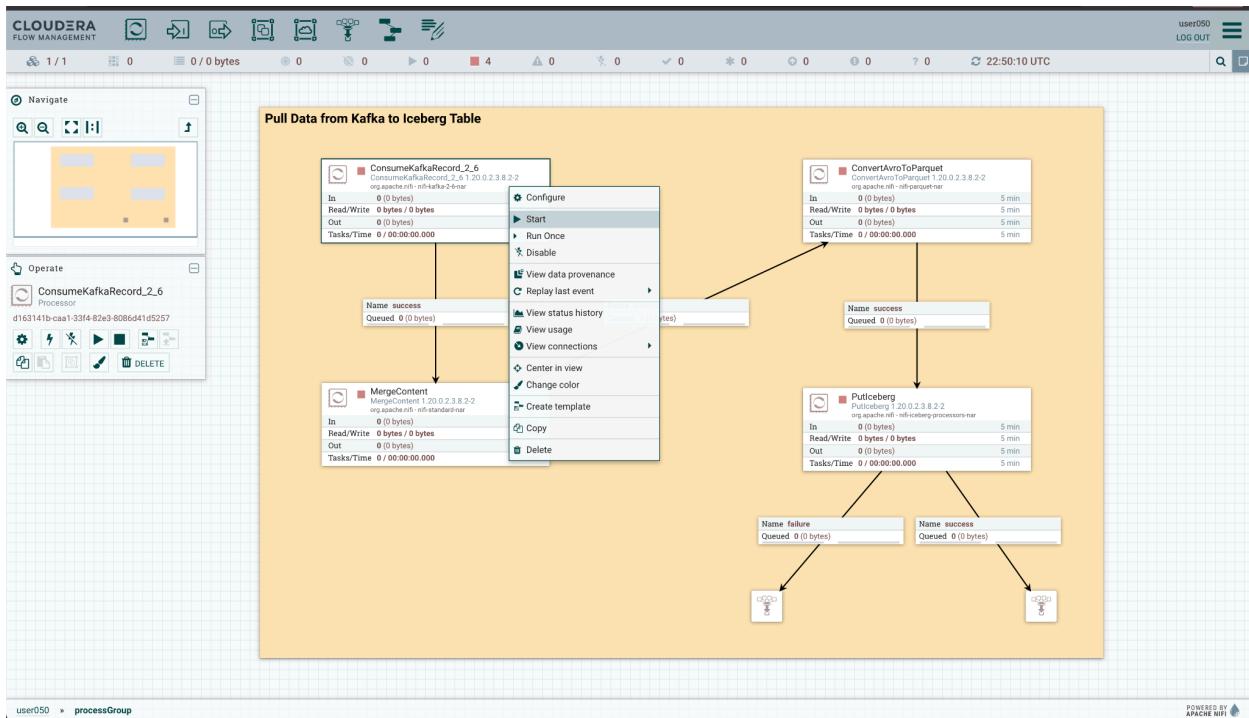
ConvertAvroToParquet, conversion needed to store the data in PARQUET format.

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.

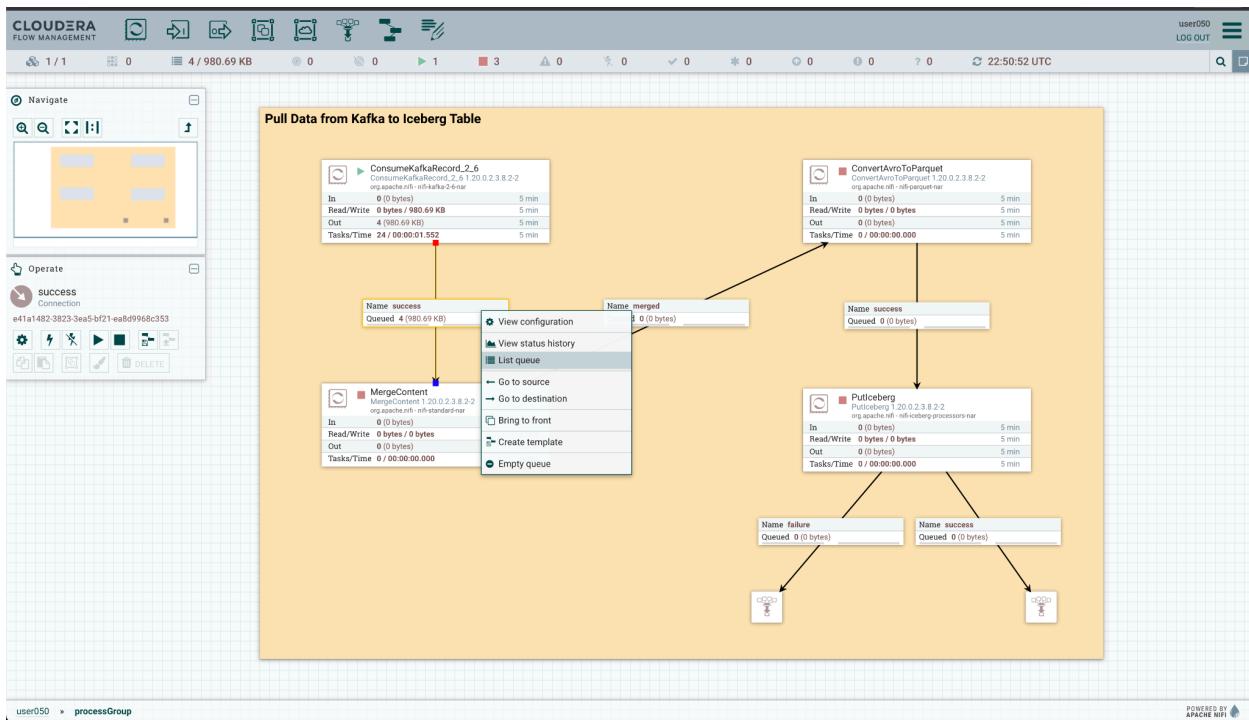


17. Now you are ready to initiate the pipeline. Start with **ConsumeKafkaRecord**, by right-clicking on it, and then clicking on **Start**. This will start consuming the Kafka topic data.



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.

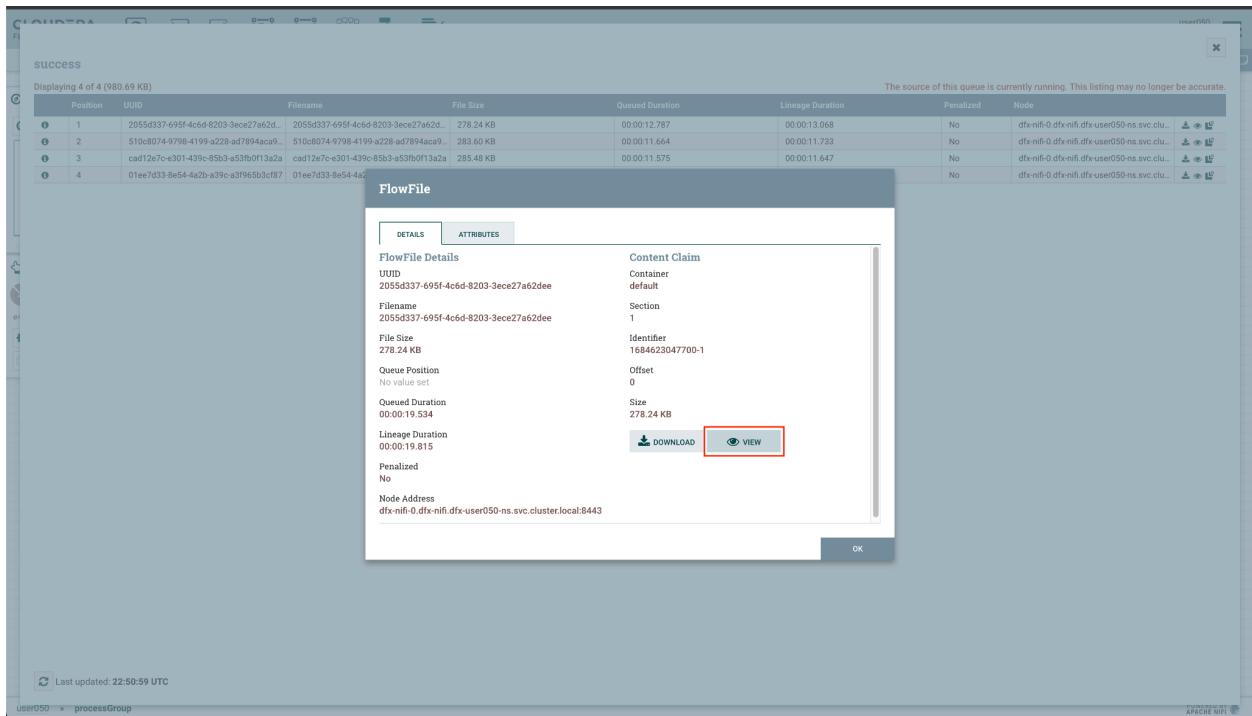


19. The next popup window lists the queued data. Click on the information icon (i) that appears on the left side to view the events.

The screenshot shows a window titled "Queue View" with the status "success". It displays a table of four queued files. The columns are: Position, UUID, Filename, File Size, Queued Duration, Lineage Duration, Penalized, and Node. A message at the top right states: "The source of this queue is currently running. This listing may no longer be accurate." Below the table, there is a note: "Last updated: 22:50:59 UTC". At the bottom, it shows the process group "user050" and the Apache NiFi logo.

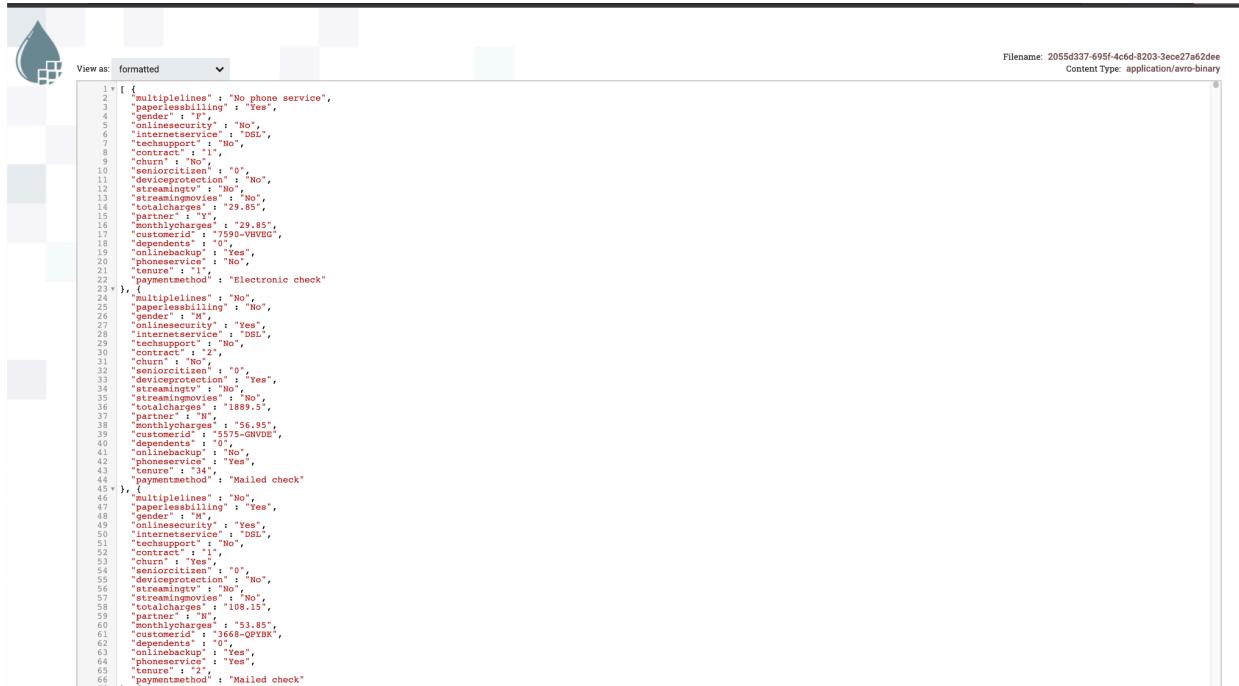
Position	UUID	Filename	File Size	Queued Duration	Lineage Duration	Penalized	Node
1	2055d337-695f-4c6d-8203-3ece27a62d...	2055d337-695f-4c6d-8203-3ece27a62d...	278.24 KB	00:00:12.787	00:00:13.068	No	dfx-nifi-0-dfx-nifi.dfx-user050-n.svc.clu...
2	510c8074-9798-4199-a228-ad7894ac9...	510c8074-9798-4199-a228-ad7894ac9...	283.60 KB	00:00:11.664	00:00:11.733	No	dfx-nifi-0-dfx-nifi.dfx-user050-n.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-n.svc.clu...
4	01ee7d33-8e54-4a2b-a39c-a3f9e5b3cf87	01ee7d33-8e54-4a2b-a39c-a3f9e5b3cf87	133.37 KB	00:00:11.527	00:00:11.567	No	dfx-nifi-0-dfx-nifi.dfx-user050-n.svc.clu...

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



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.

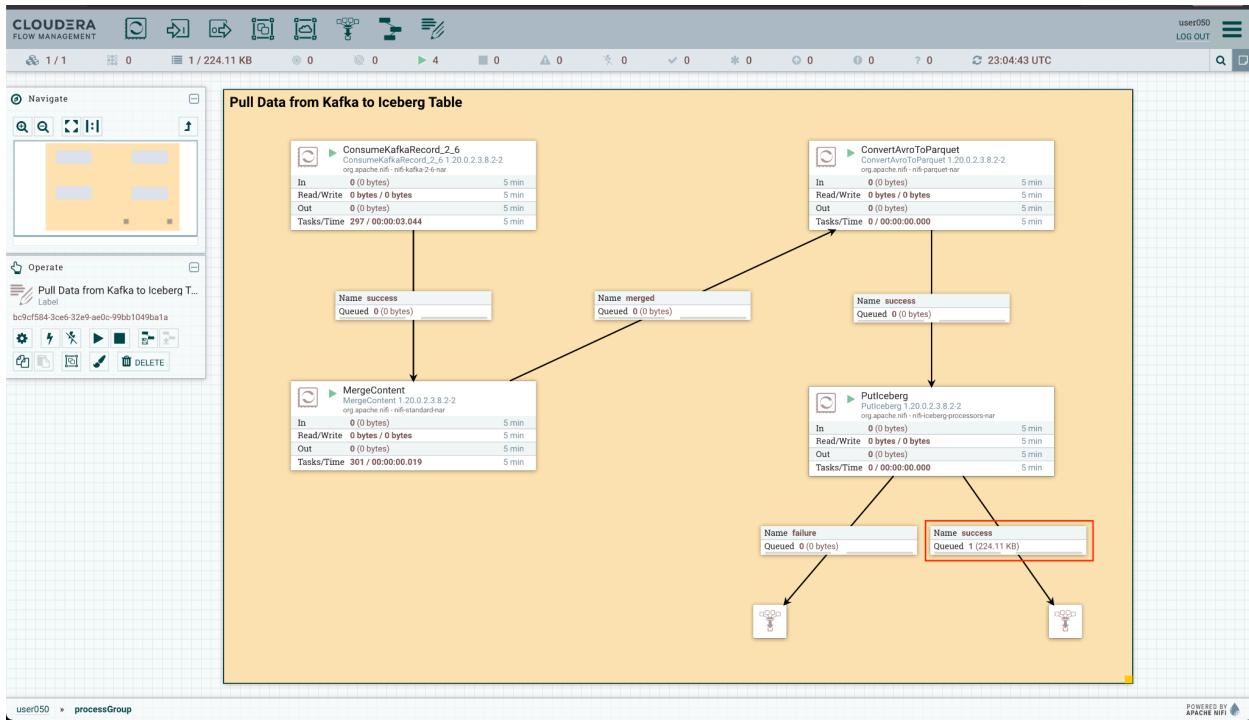


The screenshot shows a terminal window with the following details:

- Icon: A blue water droplet icon.
- Text: "View as: formatted".
- Text: "Filename: 2055d337-695f-4c6d-8203-3ece27a62dee
Content Type: application/avro-binary".
- Text: The content of the file is a JSON array with 66 elements. Each element is an object with various fields such as "multiplelines", "paperlessbilling", "onlinesecurity", "internetservice", "techsupport", "contract", "churn", "seniorcitizen", "deviceprotection", "streamingtv", "streamingmovies", "totalcharges", "monthlycharges", "customerid", "gender", "partner", "dependents", "phoneservice", "tenure", and "paymentmethod". The values for these fields are strings representing categorical or numerical data.

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**.



End of Lab 1