

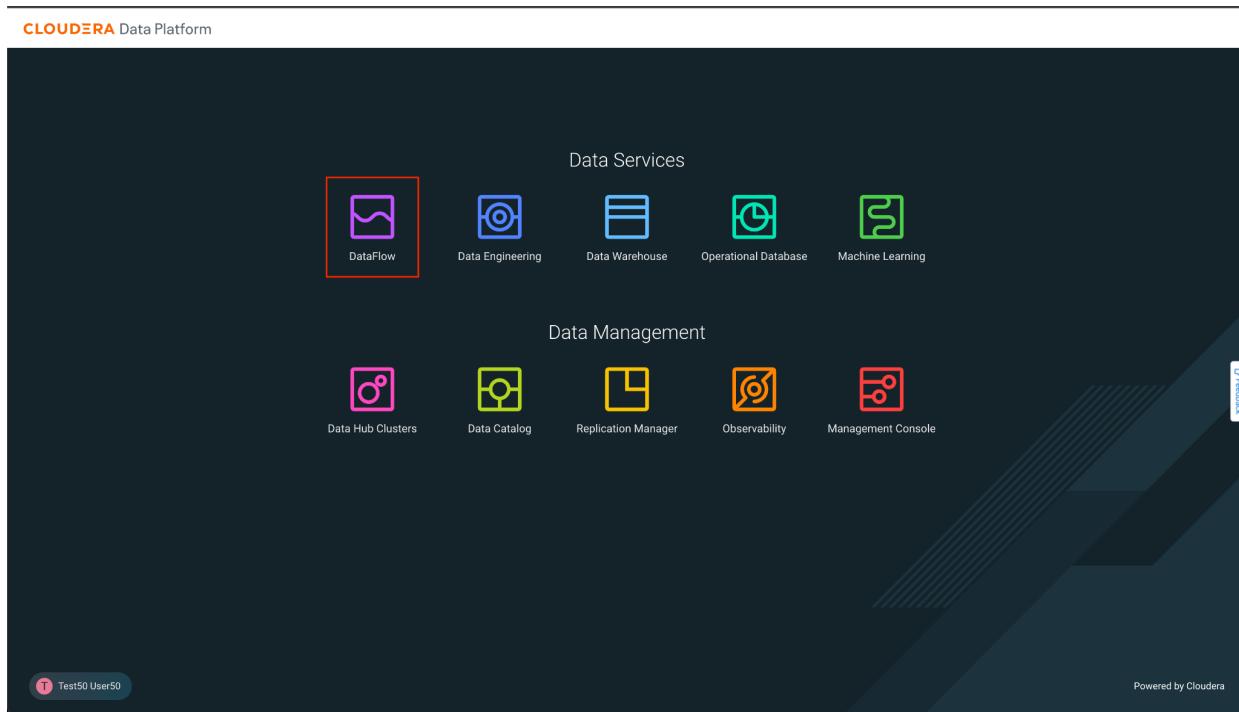
Data Lifecycle CDP Public Cloud

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:



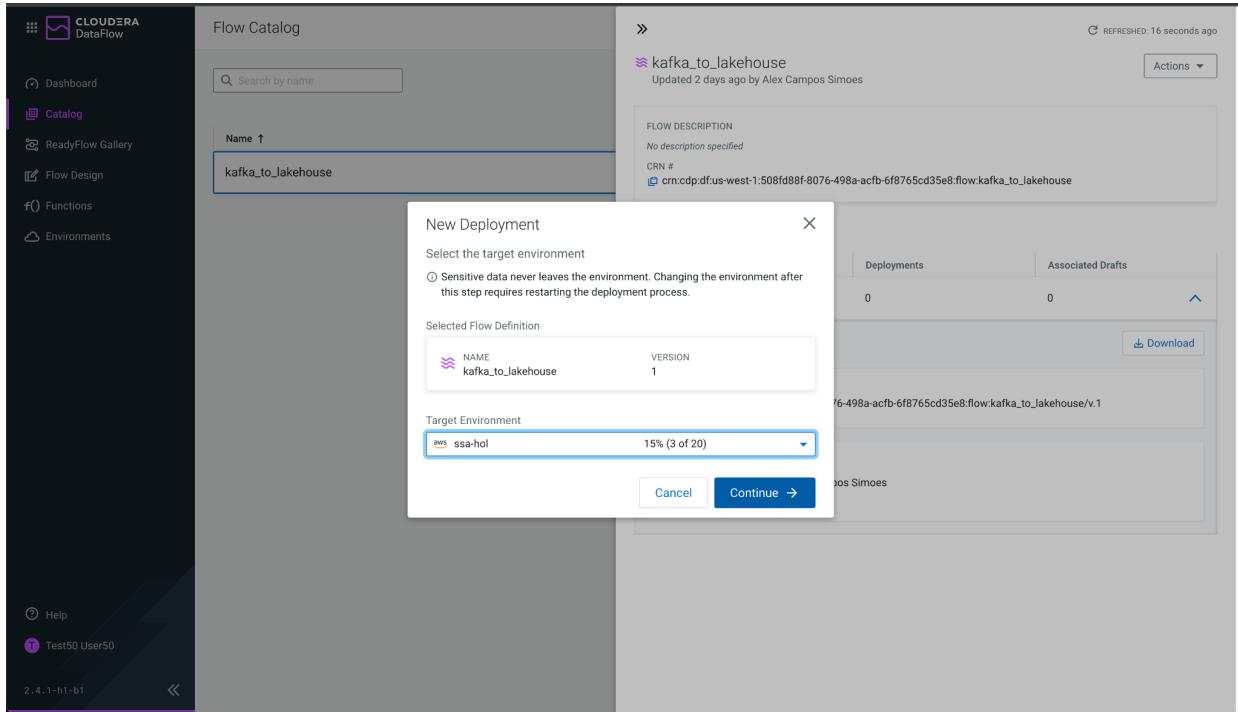
2. Once in DataFlow, click on the option **Catalog** from the left menu. The data ingestion application templates are listed here. For the purpose of this workshop, we have created and published a template that allows you to read Kafka topic data and ingest/store it in the Lakehouse provided by CDP Public 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 Test50 User50. Below the sidebar is a footer with the text "2.4.1-h1-b1". The main content area is titled "Flow Catalog" and contains a search bar labeled "Search by name". A table lists one flow entry: "kafka_to_lakehouse" (Type: Custom Flow Definition, Version: 1, Last Updated: 2 days ago). At the top right of the table are buttons for "Import Flow Definition" and "REFRESHED: 5 seconds ago". At the bottom right 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 as the previous one, but the "kafka_to_lakehouse" flow entry is now selected, indicated by a blue border around its row in the catalog table. The right-hand panel displays detailed information about this flow. At the top right is a "Actions" dropdown menu. Below it, the flow's name is shown as "kafka_to_lakehouse" with a small purple icon, followed by the text "Updated 2 days ago by Alex Campos Simoes". Under "FLOW DESCRIPTION", there is a note "No description specified" and a CRN field containing "crm.cdp.df.us-west-1:508fd88f-8076-498a-acfb-6f8765cd35e8:flow:kafka_to_lakehouse". A checkbox labeled "Only show deployed versions" is present. The main table shows one version (Version 1) with 0 deployments and 0 associated drafts. A "Deploy →" button is located at the bottom of this table. To the right of the table is a "Download" button with a cloud icon. At the bottom of the panel, under "CREATED", is the timestamp "2023-05-19 00:15 CEST by Alex Campos Simoes" and the note "'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 kafka_to_lakehouse | VERSION 1 |
|----------------------------|--------------|

Target Environment

| | |
|-----|-----------------|
| aws | NAME ssa-hol |
|-----|-----------------|

[Cancel](#) [Next →](#)

6. Uncheck the option **Automatically start flow upon successful deployment** and click **Next**.

We are going to run Flow step by step, so we don't want it to start automatically.

New Deployment

NiFi Configuration

NiFi Runtime Version

| | |
|--|----------------|
| CURRENT VERSION Latest Version (1.20.0.2.3.8.2-2) | Change Version |
|--|----------------|

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

Autostart Behavior

Automatically start flow upon successful deployment

Inbound Connections

Allow NiFi to receive data

Custom NAR Configuration

This flow deployment uses custom NARs

Overview

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

[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

Sizing & Scaling

Key Performance Indicators

Review

Parameters

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

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

SHOW: Sensitive No value

parameters (7)

CDP Workload User Password

CDP Workload Username

CDPEnvironment

core-site.xml

ssl-client.xml

hive-site.xml

Select File

Drop file or browse

Flow Definition: kafka_to_lakehouse v.1

Environment Deploying To: ssa-hd

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

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

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

1. Overview

2. NiFi Configuration

3. Parameters

4. Sizing & Scaling

5. Key Performance Indicators

6. Review

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
CDP WORKLOAD USER PASSWORD [Sensitive Value Provided]
CDP 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

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

Key Performance Indicators

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

[Learn more ↗](#)

Add New KPI

Cancel ← Previous Next →

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

New Deployment

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

Review

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

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 icon for Test50 User50. The main area is titled 'Dashboard' and shows a table of flows. A single row is selected, showing 'user050' in the 'Name' column and 'Deploying' in the 'Status' column. To the right of the table is a detailed view for 'user050'. This view includes tabs for 'KPIs', 'System Metrics', and 'Alerts'. The 'Alerts' tab is active, displaying a message: 'Deployment Initiated' followed by 'Initiated deployment of [user050]'. Below the alerts, there's a section for 'Active Alerts' which says 'No alerts to display.' There's also a 'Event History' section with a dropdown menu set to 'Deployment Initiated' and a date of '2023-05-21 00:09 CEST'. A blue button labeled 'Load More' is located at the bottom of this history section.

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

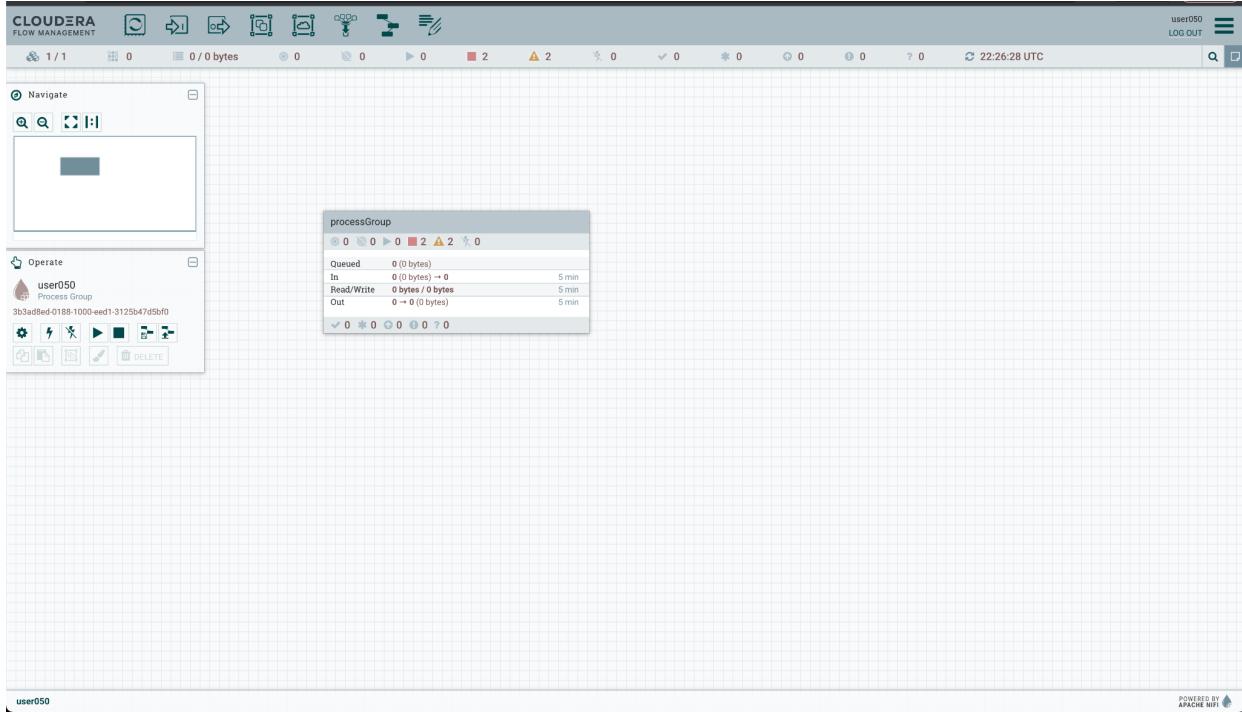
The screenshot shows the Cloudera DataFlow interface. On the left sidebar, there are links for Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, Environments, Help, and Test50 User50. The main area is titled 'Dashboard' and shows a deployment named 'user050' with status 'Deploying'. The 'Alerts' tab is active, showing 'No alerts to display.' Below it, the 'Event History' section lists various deployment 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 |

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'. It displays deployment details such as flow definition ('kafka_to_lakehouse V.1'), node count (1), environment ('aws ssa-hol'), and region ('US East(N. Virginia)'). The 'Actions' dropdown menu includes options like 'View in NiFi', 'Start flow', 'Change NiFi Runtime Version', 'Restart Deployment', and 'Terminate'.

14. In the new window you should be able to see the Flow Management canvas with one process group (a box). The canvas is where the Flow Management applications are built. Double click on the box; the only visible box, which is a Process Group and should be titled **processGroup**.



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

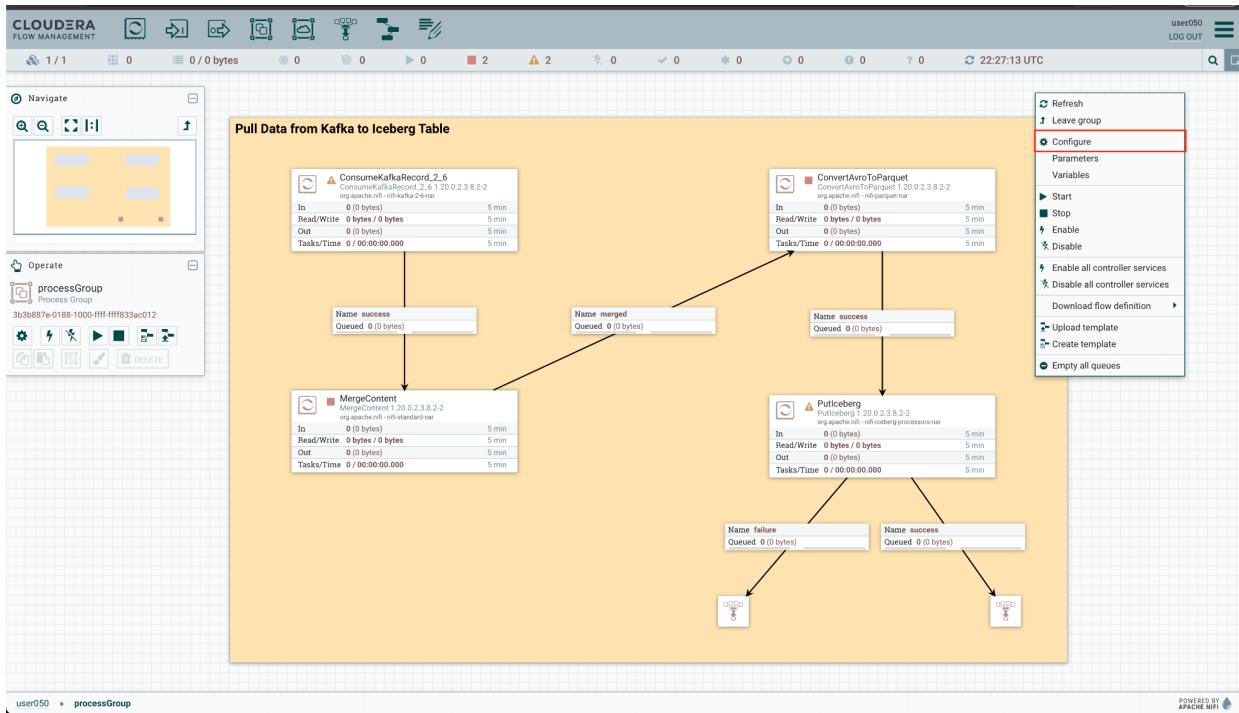
ConsumeKafkaRecord, processor to consume data from the Kafka topic, reading the data in JSON format and outputting in AVRO format.

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

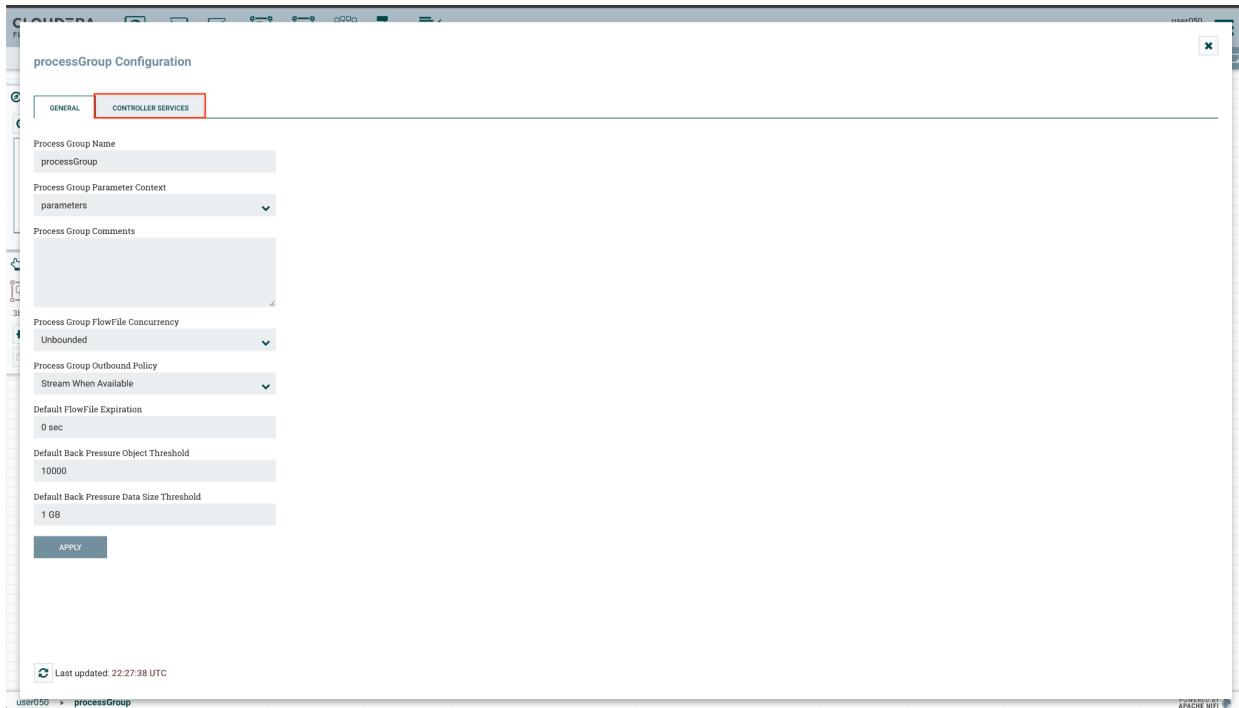
ConvertAvroToParquet, conversion needed to store the data in PARQUET format.

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, and some have an error message/alert icon. The latter is because there are components of the data flow that must be activated before. To activate them - the *Controller Services* - right click on the canvas and click on the option **Configure** from the floating menu that appears.

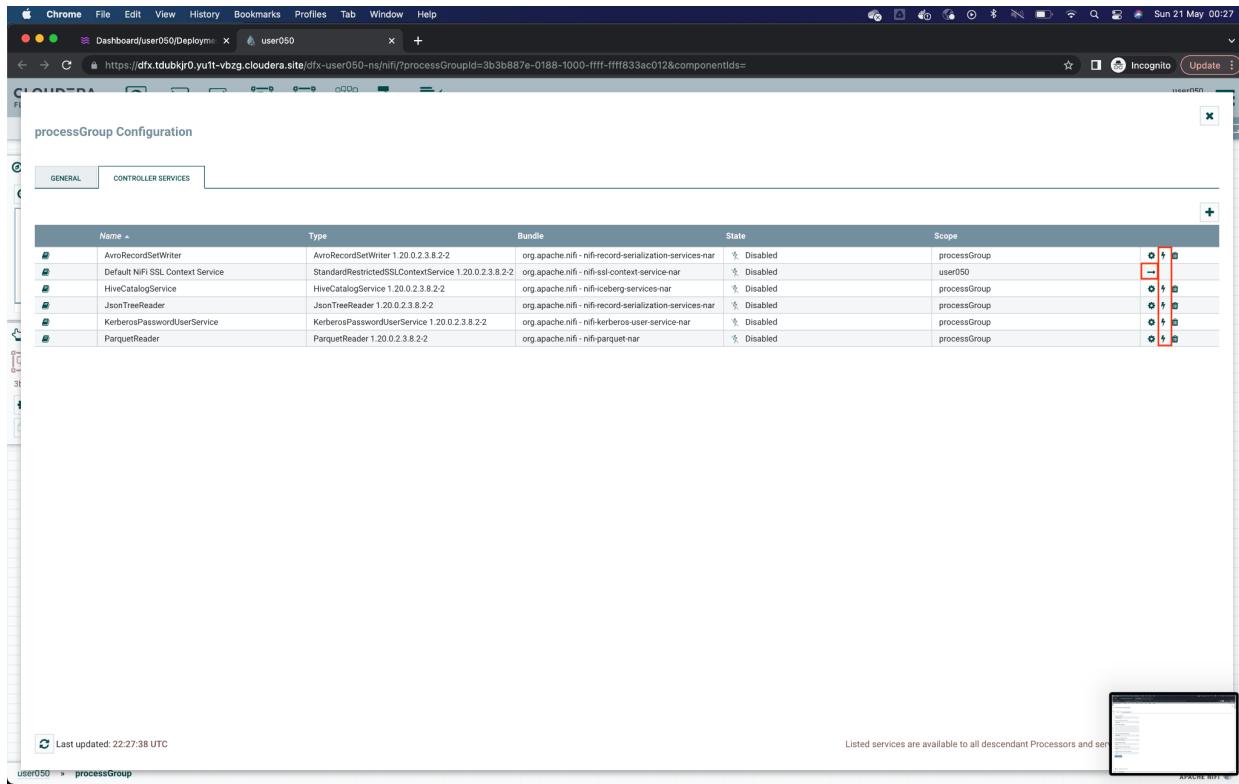


16. In the pop-up window that opens, select the tab **Controller Services**.



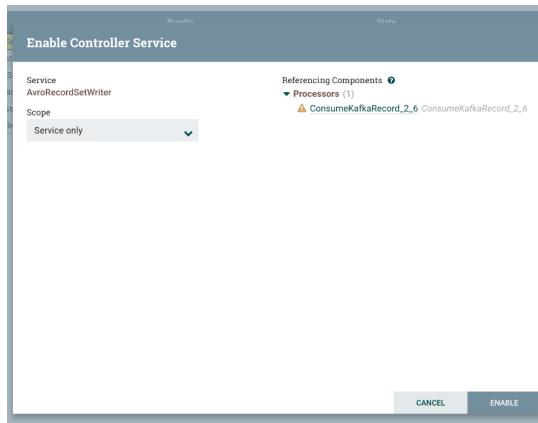
17. The **Controller Services** of the data flow. Each of them must be activated. The following Controllers must be activated first: **AvroReaderSetWriter**, **HiveCatalogService**, **JsonTreeReader**, **KerberosPasswordUserService** and **ParquetReader** clicking on the icon

lightning  which appears on the right (marked in red).

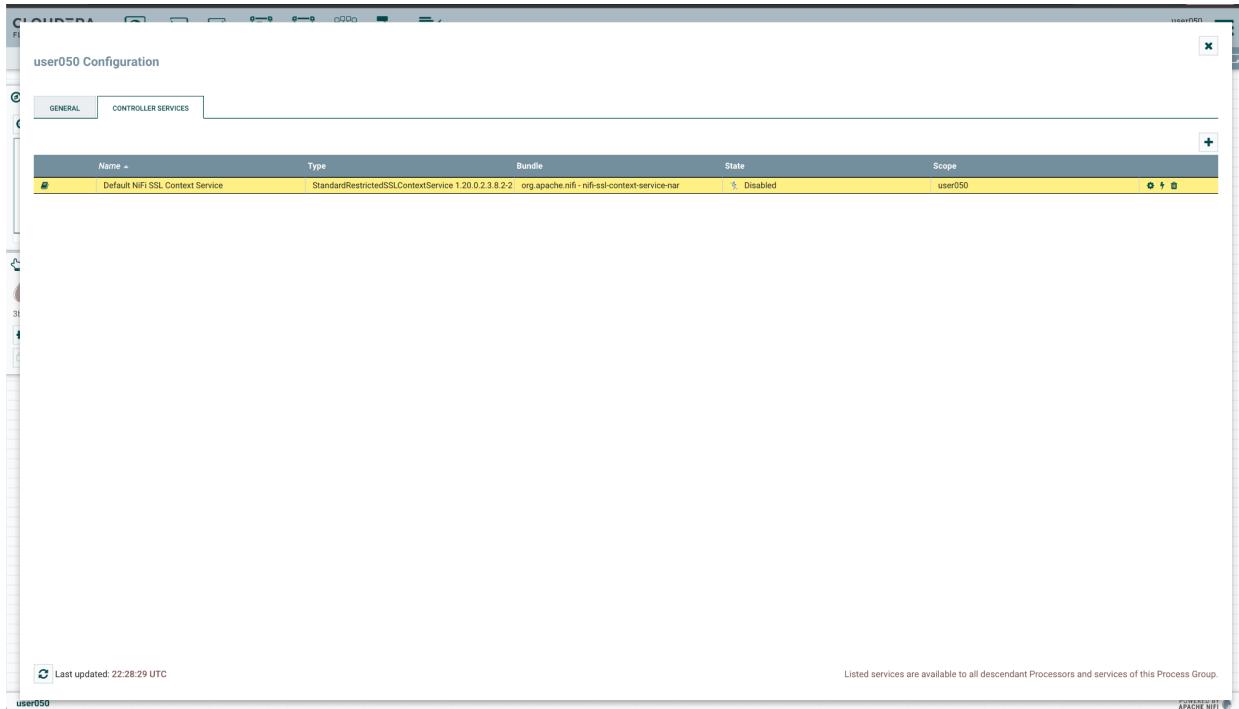


| Name | Type | Bundle | State | Scope |
|----------------------------------|--|--|----------|--------------|
| AvroRecordSetWriter | AvroRecordSetWriter 1.20.0.2.3.8.2-2 | org.apache.nifi - nifi-record-serialization-services-nar | Disabled | processGroup |
| Default NiFi SSL Context Service | StandardRestrictedSSLContextService 1.20.0.2.3.8.2-2 | org.apache.nifi - nifi-ssl-context-service-nar | Disabled | user050 |
| HiveCatalogService | HiveCatalogService 1.20.0.2.3.8.2-2 | org.apache.nifi - nifi-catalog-services-nar | Disabled | processGroup |
| JsonTreeReader | JsonTreeReader 1.20.0.2.3.8.2-2 | org.apache.nifi - nifi-record-serialization-services-nar | Disabled | processGroup |
| KerberosPasswordEncoderService | KerberosPasswordEncoderService 1.20.0.2.3.8.2-2 | org.apache.nifi - nifi-kerberos-user-service-nar | Disabled | processGroup |
| ParquetReader | ParquetReader 1.20.0.2.3.8.2-2 | org.apache.nifi - nifi-parquet-nar | Disabled | processGroup |

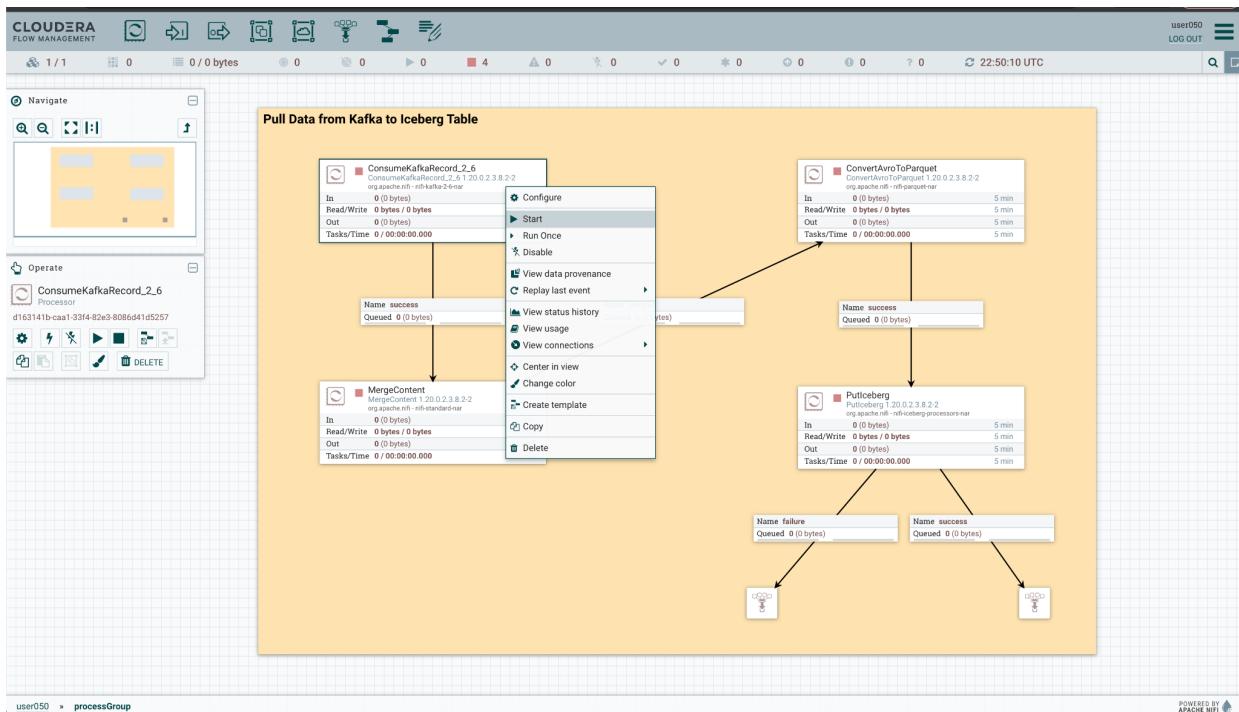
Click the button **Enable** in the enable confirmation window of each Controller Service. Then close that window to enable the next Controller Service.



To activate the Default NiFi SSL Context Service, you must click on the arrow  . Finally clicking on the lightning bolt icon  controller service is activated **Default NiFi SSL Context Service**, which will also present a window to enable it.

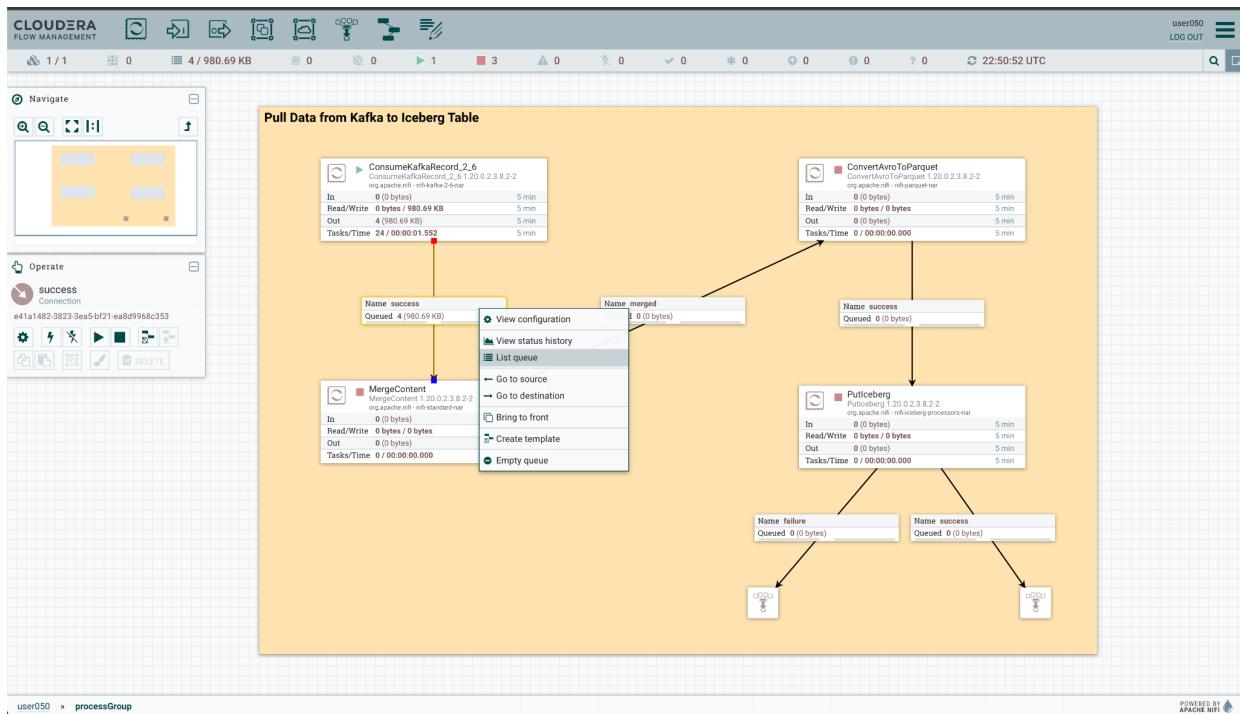


18. Close the Controller Services window, making sure all are enabled. Return to the Process Group by double-clicking on it. It's time to execute **Processors**. Start with **ConsumeKafkaRecord**, by right-clicking on it, and then clicking on **Start**. This will start consuming the Kafka topic data.



19. 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.



20. 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-ad7894aca9... | 510c8074-9798-4199-a228-ad7894aca9... | 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." At the bottom left, it says "Last updated: 22:50:59 UTC". At the bottom right, it says "APACHE NIFI".

21. 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 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 the same as in the previous screenshot. Below the table, a message says "The source of this queue is currently running. This listing may no longer be accurate." At the bottom left, it says "Last updated: 22:50:59 UTC". At the bottom right, it says "APACHE NIFI".

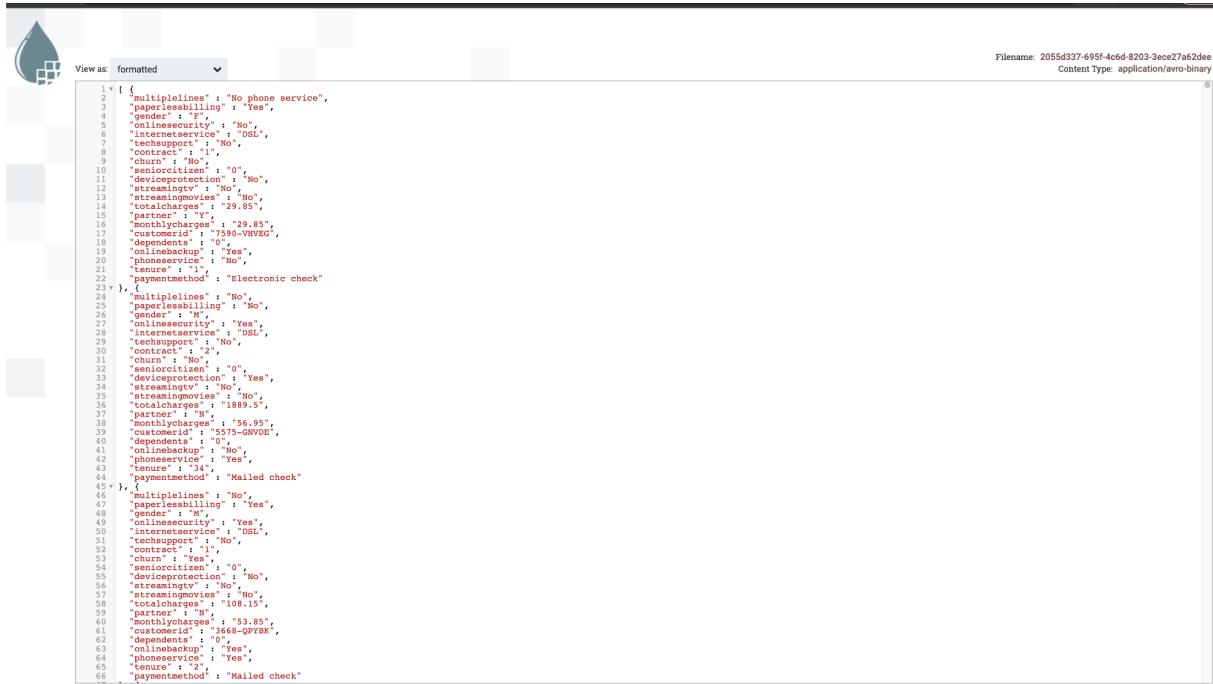
A modal dialog box is open, showing "FlowFile Details". It has two tabs: "DETAILS" (selected) and "ATTRIBUTES". The "DETAILS" tab displays the following information:

| 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 | VIEW |
| Penalized No | |
| Node Address dfx-nifi-0.dfx-nifi.dfx-user050.ns.svc.cluster.local:8443 | |

The "VIEW" button is highlighted with a red box. At the bottom right of the dialog, there is an "OK" button.

22. 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**.

23. 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 file viewer interface with a "View as" dropdown set to "formatted". The content is an Avro binary file containing two customer records. Record 1 (customerid: 100-VWEG) has a single-line contract, a monthly charge of 29.85, and a total charge of 29.85. Record 2 (customerid: 5575-GNVD8) has a two-line contract, a monthly charge of 56.95, and a total charge of 1889.51. Both records include fields like multipiplelines, paperlessbilling, onlinesecurity, internetservice, contract, churn, seniorcitizen, deviceprotection, streamingmovies, totalcharges, monthlycharges, phonelines, and paymentmethod.

```

1+ [{"multipiplelines": "No phone service",
2  "paperlessbilling": "Yes",
3  "seniorcitizen": "0",
4  "deviceprotection": "No",
5  "streamingmovies": "No",
6  "internetservice": "DSL",
7  "phonelines": "No",
8  "contract": "1",
9  "churn": "0",
10 "seniorcitizen": "0",
11 "deviceprotection": "No",
12 "streamingmovies": "No",
13 "totalcharges": "29.85",
14 "monthlycharges": "29.85",
15 "customerid": "100-VWEG",
16 "paymentmethod": "Electronic check"
17 }, {"multipiplelines": "No",
18 "paperlessbilling": "No",
19 "onlinesecurity": "Yes",
20 "internetservice": "DSL",
21 "phonelines": "Yes",
22 "contract": "2",
23 "churn": "0",
24 "seniorcitizen": "0",
25 "deviceprotection": "Yes",
26 "streamingmovies": "No",
27 "totalcharges": "1889.51",
28 "monthlycharges": "56.95",
29 "customerid": "5575-GNVD8",
30 "phonelines": "Yes",
31 "onlinebackup": "No",
32 "phoneservice": "Yes",
33 "tenure": "34",
34 "paymentmethod": "Mailed check"
35 }, {"multipiplelines": "No",
36 "paperlessbilling": "Yes",
37 "onlinesecurity": "Yes",
38 "internetservice": "DSL",
39 "phonelines": "No",
40 "contract": "1",
41 "churn": "Yes",
42 "seniorcitizen": "0",
43 "deviceprotection": "No",
44 "streamingmovies": "No",
45 "totalcharges": "108.15",
46 "monthlycharges": "53.85",
47 "customerid": "3668-QPVK8",
48 "phonelines": "Yes",
49 "onlinebackup": "Yes",
50 "phoneservice": "Yes",
51 "tenure": "2",
52 "paymentmethod": "Mailed check"
53 }, {"multipiplelines": "No",
54 "paperlessbilling": "Yes",
55 "onlinesecurity": "Yes",
56 "internetservice": "No",
57 "phonelines": "Yes",
58 "contract": "1",
59 "churn": "Yes",
60 "seniorcitizen": "0",
61 "deviceprotection": "No",
62 "streamingmovies": "Yes",
63 "totalcharges": "29.85",
64 "monthlycharges": "29.85",
65 "customerid": "100-VWEG",
66 "paymentmethod": "Electronic check"}]

```

24. 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**.

