

CLOUDERA

DATA FLOW HANDS - ON

STUDENT GUIDE

Pre-requisites	5
Lab 0 - Introduction and setup	6
1. Verify access to the workshop environment	6
2. Verify permissions in Apache Ranger	7
2.1 Accessing Apache Ranger	7
2.2 Kafka Permissions	8
2.3 Schema Registry Permissions	10
3. Update workload password	11
4. Obtain the Kafka Broker List	13
Step 1 : Access the Data Hub	13
Step 2 : Go to the Streams Messaging Interface	14
Step 3 : Select Brokers from the left tab	14
Step 4 : Save the broker list	15
5. Download Resources from GitHub	16
Step 1 : Access the URL shared by the instructor for GitHub	16
Step 2 : Download the repo as a zip file	16
Step 3 : Uncompress the Files	16
6. Unlock your KeyTab	17
1. Unlock your Keytab if it is not unlocked already	17
Step 1 : Go to the SSB Data Hub	17
Step 2 : Open the SSB UI by clicking on Streaming SQL Console	18
Step 3 : Click on the User name at the bottom left of the screen and select Manage Keytab	18
Step 4 : Enter your Workload Username (apacXY) and Password.	19
Step 5 : Click on unlock KeyTab	19
2. Reset your KeyTab if it is already unlocked	20
Step 1 : Go to the SSB Data Hub	20
Step 2 : Open the SSB UI by clicking on Streaming SQL Console	21
Step 3 : Click on the User name at the bottom left of the screen and select Manage Keytab	21
Lab 1 : Create a Flow using the Flow Designer	23
1. Overview	23
2. Building the Data Flow	23
2.1. Create the canvas to design your flow	23
Step 1: Access the DataFlow Data Service	23
Step 2: Go to the Flow Design	24
Step 3: Create a new Draft	24
Step 4: Select the appropriate environment	24
2.2. Adding new parameters	26
Step 1: Click on the FLOW OPTIONS on the top right corner of your canvas and then	

Step 2: Configure Parameters	26
2.3. Create the Flow	29
Step 1: Add GenerateFlowFile processor	30
Step 2: Configure GenerateFlowFile processor	31
Step 3: Add PutCDPObjectStore processor	34
Step 4: Configure PutCDPObjectStore processor	35
Step 5: Create connection between processors	37
2.4. Naming the queues	39
3. Testing the Data Flow	41
Step 1: Start test session	41
Step 2: Run the flow	42
4. Move the Flow to the Flow Catalog	44
Step 1: STOP the current test session	44
Step 2: PUBLISH the flow	45
Step 3: Give your flow a name and click on PUBLISH	46
5. Deploying the Flow	47
Step 1: Search for the flow in the Flow Catalog	47
Step 2: Deploy the flow	48
Step 3: Select the CDP environment	48
Step 4: Deployment Name	49
Step 5: Set the NiFi Configuration	49
Step 6: Set the Parameters	50
Step 7: Set the cluster size	50
Step 8: Add Key Performance indicators	51
Step 9: Click Deploy	53
6. Viewing details of the deployed flow	54
Step 1 : Manage KPI and Alerts	54
Step 2 : Manage Sizing and Scaling	55
Step 3 : Manage Parameters	56
Step 4 : NiFi Configurations	56
Step 5 : View the deployed flow in NiFi	57
Step 6 : Terminate the flow	58
Lab 2 : Migrating Existing Data Flows to CDF-PC	60
1. Overview	60
2. Pre-requisites	61
2.1. Create a Kafka Topic	61
2.2. Create a Schema in Schema Registry	63
Lab 3 : Operationalizing Externally Developed Data Flows with CDF-PC	67
1. Import the Flow into the CDF-PC Catalog	67
2. Deploy the Flow in CDF-PC	68
Lab 4 : SQL Stream Builder	76
1. Overview	76
2. Creating a Project	76
Step 1: Go to the SQL Stream Builder UI	76

Step 3 : Create Kafka Data Store	78
Step 4: Create Kafka Table	79
Step 5: Configure the Kafka Table	80
Step 6: Create a Flink Job	83

Pre-requisites

For the ease of carrying out the workshop and considering the time at hand, we have already taken care of some of the steps that need to be considered before we can start with the actual Lab steps. The prerequisites that need to be in place are:

1. Streams Messaging Data Hub Cluster should be created and running.
2. Stream analytics Data Hub cluster should be created and running.
3. Data provider should be configured in SQL Stream Builder.
4. Have access to the file syslog-to-kafka.json.
5. Environment should be enabled as part of the CDF Data Service.

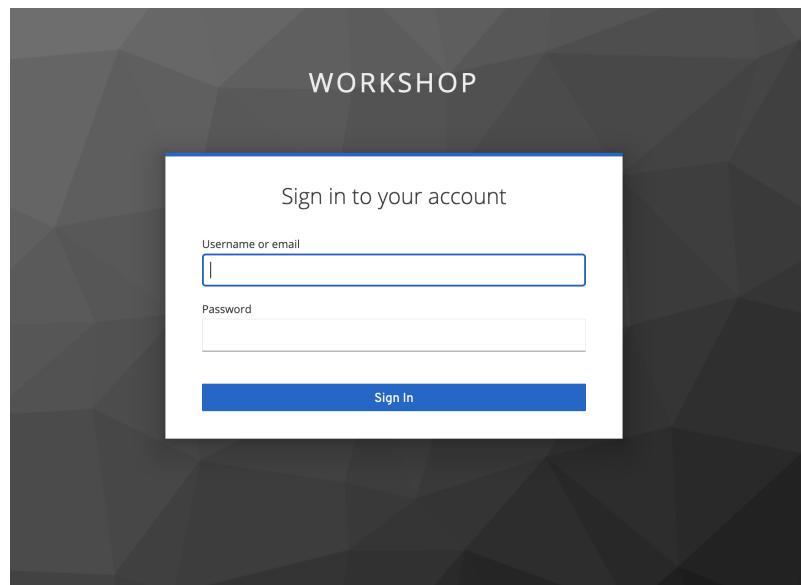
Lab 0 basically talks about verifying different aspects wrt to access and connections before we could begin with the actual steps.

Lab 0 - Introduction and setup

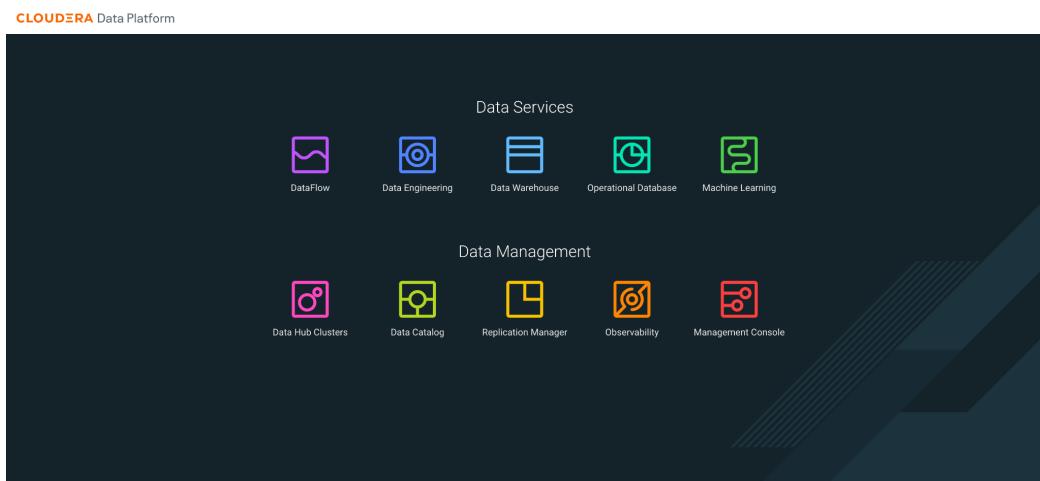
1. Verify access to the workshop environment

- The **INSTRUCTOR** will share the Workshop link and the credentials before the start of the workshop
- Open the shared link and login with the credentials assigned to you.

<Will be shared by the instructor at the start>



- You should land on the CDP Console as shown below.

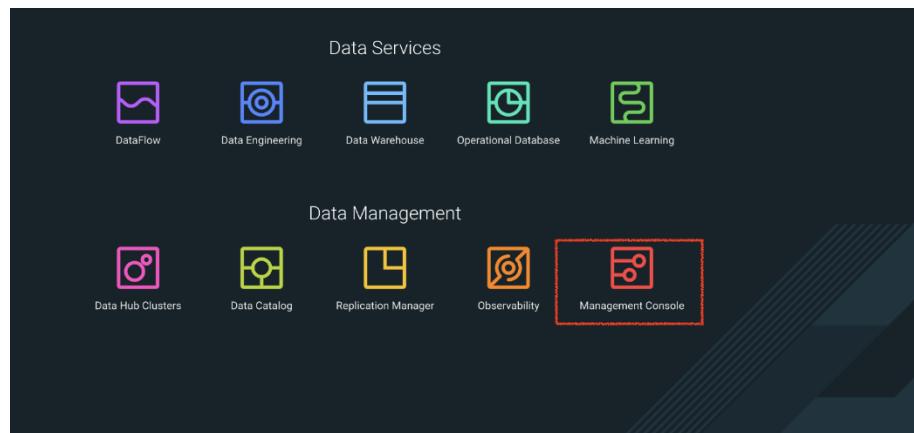


2. Verify permissions in Apache Ranger

NOTE: THESE STEPS HAVE ALREADY BEEN DONE FOR YOU, THIS SECTION WILL WALK YOU THROUGH HOW PERMISSIONS/POLICIES ARE MANAGED IN RANGER.
PLEASE DO NOT EXECUTE THE STEPS IN THIS SECTION OR CHANGE ANYTHING.

2.1 Accessing Apache Ranger

Step 1 : Click on Management Console



Step 2 : Click on Environments on the left tab

The screenshot shows the 'Environments / List' page. On the left, a sidebar menu includes 'Dashboard', 'Environments' (which is selected and highlighted in red), 'Data Lakes', 'User Management', 'Data Hub Clusters', 'Data Warehouses', 'ML Workspaces', 'Classic Clusters', 'Audit', 'Shared Resources', and 'Global Settings'. The main content area displays a table of environments:

Status	Name	Cloud Provider	Region	Data Lake	CDP Runtime Version	Time Created
Available	emeaworkshop-env	aes	Asia Pacific (Mumbai)	Running	7.2.16	4/12/2023, 4:40:07 PM GMT+5:30
Available	dss-workshop-env	aes	US East(N. Virginia)	Running	7.2.16	4/10/2023, 8:37:03 PM GMT+5:30
Environment Stopped	cml-adb	aes	Asia Pacific (Mumbai)	Stopped	7.2.16	4/4/2023, 5:20:56 PM GMT+5:30
Available	pko-hands-on-workshop-env	aes	Asia Pacific (Mumbai)	Running	7.2.16	3/28/2023, 12:24:09 PM GMT+5:30
Environment Stopped	pse-lsv-env	aes	US West (Oregon)	Stopped	7.2.16	3/23/2023, 1:22:31 AM GMT+5:30
Environment Stopped	meta-workshop	aes	EU (London)	Stopped	7.2.16	2/24/2023, 10:30:31 PM GMT+5:30
Environment Stopped	vrayker-cdp2	aes	Asia Pacific (Sydney)	Stopped	7.2.15	10/13/2022, 5:13:00 PM GMT+5:30
Environment Stopped	pse-workshop	aes	US East (Ohio)	Stopped	7.2.16	9/16/2022, 12:32:53 AM GMT+5:30

At the bottom right of the table, there are pagination controls: '1 - 8 of 8', '< >', 'Items per page: 25', and a dropdown menu.

Step 3 : Select the environment that is shared by the instructor and click on the **Ranger** quick link to access the Ranger UI

2.2 Kafka Permissions

1. In Ranger, select the Kafka repository that's associated with the stream messaging datahub.

2. Verify if the user group(**workshop-users**) who will be performing the workshop is present in both **all-consumergroup** and **all-topic**.

Policy ID	Policy Name	Policy Labels	Status	Audit Logging	Roles	Groups	Users	Action
103	all - consumergroup	--	Enabled	Enabled	--	__c_ranger_admins_3e797a6	cruisecontrol ssb streamsmgr kafka kafka_mirror_maker streamsrepmgr rangerlookup - Less...	
105	all - topic	--	Enabled	Enabled	--	__c_ranger_admins_3e797a6	cruisecontrol ssb streamsmgr kafka kafka_mirror_maker streamsrepmgr rangerlookup - Less...	

- All-consumergroup

Allow Conditions:

Select Role	Select Group	Select User	Policy Conditions	Permissions	Delegate Admin
Select Roles	x __c_ranger_admins_6276838c x workshop-users	x cruisecontrol x ssb x streamsmgr x kafka x kafka_mirror_maker x streamsrepmgr	Add Conditions + add	Consume Describe Delete	<input checked="" type="checkbox"/>
Select Roles	Select Groups	x rangerlookup	Add Conditions +	Describe	

- all-topic

Policy Details:

Policy Type	Access	Policy Conditions	Add Validity Period
Policy ID	105	No Conditions	
Policy Name *	all - topic		
Policy Label	Policy Label		
Description	topic		
Audit Logging	<input checked="" type="radio"/> Yes		

Allow Conditions:

Select Role	Select Group	Select User	Policy Conditions	Permissions	Delegate Admin
Select Roles	x __c_ranger_admins_6276838c x workshop-users	x cruisecontrol x ssb x streamsmgr x kafka x kafka_mirror_maker	Add Conditions +	Polish Consume Configure Describe Create Delete Describe Config After Config	<input checked="" type="checkbox"/>

2.3 Schema Registry Permissions

1. In Ranger, select the Schema Registry repository that's associated with the stream messaging datahub.



2. Verify if the user group(**workshop-users**) who will be performing the workshop is present in the Policy : **all - schema-group, schema-metadata, schema-branch, schema-version**.

List of Policies : kafka_smm_cluster_schemaregistry								
Policy ID	Policy Name	Policy Labels	Status	Audit Logging	Roles	Groups	Users	Action
157	all - export-import	--	Enabled	Enabled	--	c_ranger_admins_627883bc	sbb [streamsmgr] kafka schemaregistry + More...	
160	all - serde	--	Enabled	Enabled	--	c_ranger_admins_627883bc	sbb [streamsmgr] kafka schemaregistry + More...	
163	all - schema-group, schema-metadata	--	Enabled	Enabled	--	c_ranger_admins_627883bc	sbb [streamsmgr] kafka schemaregistry + More...	
165	all - schema-group, schema-metadata, sch...	--	Enabled	Enabled	--	c_ranger_admins_627883bc	sbb [streamsmgr] kafka schemaregistry + More...	
167	all - registry-service	--	Enabled	Enabled	--	c_ranger_admins_627883bc	sbb [streamsmgr] kafka schemaregistry + More...	
169	all - schema-group, schema-metadata, sch...	--	Enabled	Enabled	--	c_ranger_admins_627883bc workshop-users	sbb [streamsmgr] kafka schemaregistry + More...	

Policy Details:

Policy Type	Access
Policy ID	101
Policy Name *	all - schema-group, schema-metadata, schema-t
Policy Label	Policy Label
schema-grp *	[x *] <input type="text"/> <input type="button" value="Include"/>
Schema Name *	[x *] <input type="text"/> <input type="button" value="Include"/>
schema-brn *	[x *] <input type="text"/> <input type="button" value="Include"/>
schema-ver *	[x *] <input type="text"/> <input type="button" value="Include"/>
Description	Policy for all - schema-group, schema-metadata, schema-branch, schema-version
Audit Logging	Yes

Allow Conditions:

Select Role	Select Group	Select User	Policy Conditions	Permissions	Delegate Admin
[Select Roles]	[x_c_ranger_admin_627683bc] [x workshop-users] [x sb] [x streamsmgr] [x kafka] [x schemaregistry] [x rangerlookup]	[x_c_ranger_admins_6021359e] [x workshop-aesean-group1]	Add Conditions +	Create Read Update Delete	<input checked="" type="checkbox"/> <input type="button" value="X"/>

Lastly, add the user group to the Streaming Analytics Data Hub Kafka Cluster Ranger permission as below:

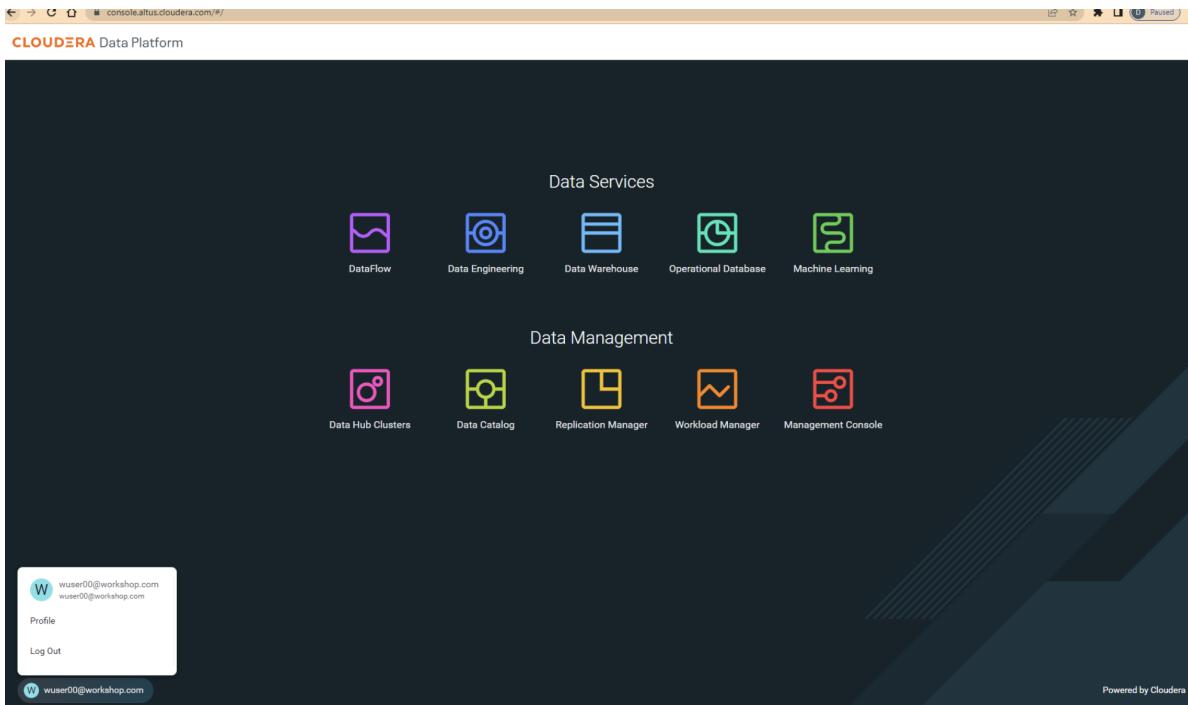
118	all - cluster	--	Enabled	Enabled	--	c_ranger_admins_6021359e workshop-aesean-group1	cruisecontrol sb streamsmgr kafka + More..	<input type="button" value="Edit"/> <input type="button" value="Delete"/>
Ranger Access Manager Audit Security Zone Settings asoni								
Service Manager > workshop_aesean_stream_analytics_kafka_e2c3 ... Last Response Time : 05/06/2023 02:59:25								
List of Policies : workshop_aesean_stream_analytics_kafka_e2c3								
<input type="text" value="Search for your policy..."/> <input type="button" value="Add New Policy"/>								
Policy ID	Policy Name	Policy Labels	Status	Audit Logging	Roles	Groups	Users	Action
115	all - consumergroup	--	Enabled	Enabled	--	c_ranger_admins_6021359e workshop-aesean-group1	cruisecontrol sb streamsmgr kafka + More..	<input type="button" value="Edit"/> <input type="button" value="Delete"/>
116	all - topic	--	Enabled	Enabled	--	c_ranger_admins_6021359e workshop-aesean-group1	cruisecontrol sb streamsmgr kafka + More..	<input type="button" value="Edit"/> <input type="button" value="Delete"/>
117	all - transactionalid	--	Enabled	Enabled	--	c_ranger_admins_6021359e	cruisecontrol sb streamsmgr kafka + More..	<input type="button" value="Edit"/> <input type="button" value="Delete"/>
118	all - cluster	--	Enabled	Enabled	--	c_ranger_admins_6021359e workshop-aesean-group1	cruisecontrol sb streamsmgr kafka + More..	<input type="button" value="Edit"/> <input type="button" value="Delete"/>
119	all - delegationtoken	--	Enabled	Enabled	--	c_ranger_admins_6021359e	cruisecontrol sb streamsmgr kafka + More..	<input type="button" value="Edit"/> <input type="button" value="Delete"/>
120	connect internal - topic	--	Enabled	Enabled	--		kafka streamsmgr	<input type="button" value="Edit"/> <input type="button" value="Delete"/>

3. UPDATE WORKLOAD PASSWORD

NOTE: THESE STEPS NEED TO BE PERFORMED BEFORE MOVING FORWARD

You will need to define your CDP Workload Password that will be used to access non-SSO interfaces. You may read more about it here. Please keep it with you. If you have forgotten it, you will be able to repeat this process and define another one.

- Click on your user name (Ex: apac00@workshop.com) at the lower left corner.
Click on **Profile**.



- Click option **Set Workload Password**.

The screenshot shows the Cloudera Management Console interface. On the left is a dark sidebar with navigation links like Dashboard, Environments, Data Lakes, User Management, Data Hub Clusters, Data Warehouses, ML Workspaces, Classic Clusters, Audit, Shared Resources, and Global Settings. The main content area has a header 'Users / wuser00@workshop.com'. Below the header, there's a table with columns Name, Email, Workload User Name, CRN, Tenant ID, Identity Provider, Last Interactive Login, and Profile Management. The 'Workload Password' row has a 'Set Workload Password' button, which is highlighted with a red box. At the bottom of the main area, there are tabs for Access Keys, Roles, Resources, Groups, and SSH Keys. A message 'No access keys found.' is displayed above the 'Generate Access Key' button.

- Enter the shared password.

NOTE: PLEASE ENTER THE SAME PASSWORD THAT WAS SHARED BY THE INSTRUCTOR. FAILING TO DO SO WILL LEAD TO ERRORS IN OUR LAB STEPS LATER ON

The screenshot shows the 'Workload Password' configuration page. It has fields for 'Password' and 'Confirm Password', both containing '*****'. Below the fields is a note: 'If you use keytabs, you need to regenerate them after changing your workload password. You can do this from your user profile > Actions > Get Keytab.' At the bottom is a blue 'Set Workload Password' button, which is highlighted with a red box.

- Click the button Set Workload Password.

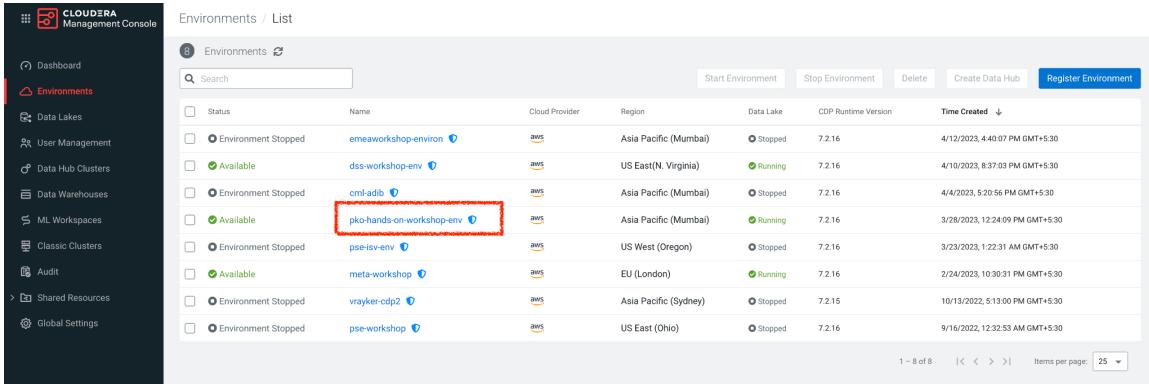
4. Obtain the Kafka Broker List

We will require the broker list to configure our processors to connect to our Kafka brokers which allow consumers to connect and fetch messages by partition, topic or offset.

This information can be found in the Data Hub cluster associated to the Streams Messaging Manager

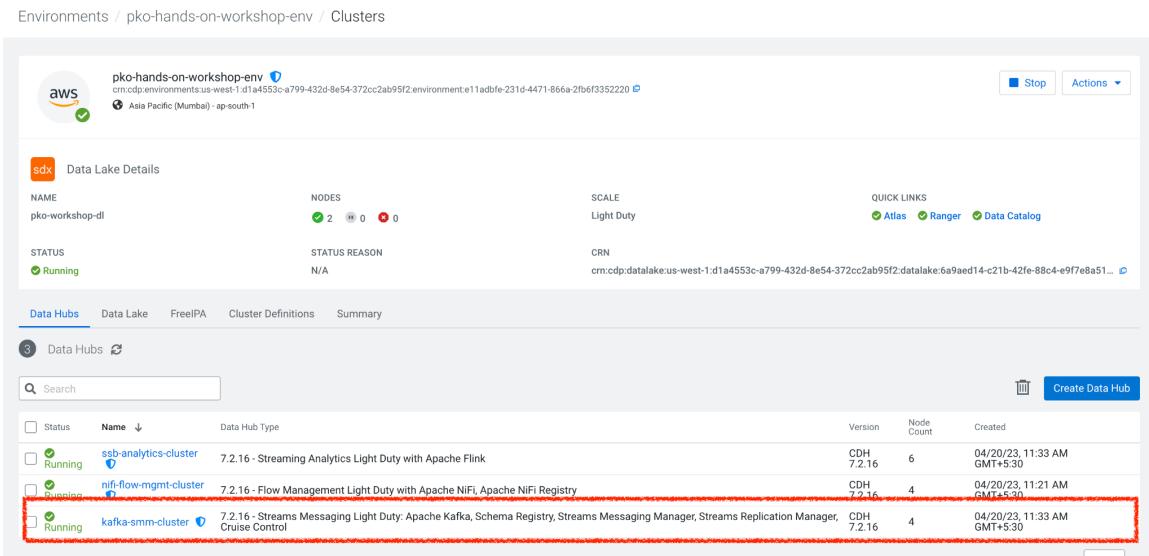
Step 1 : Access the Data Hub

- Go to the environment that is shared by the INSTRUCTOR



Status	Name	Cloud Provider	Region	Data Lake	CDP Runtime Version	Time Created
Environment Stopped	emeaworkshop-env	aws	Asia Pacific (Mumbai)	Stopped	7.2.16	4/12/2023, 4:40:07 PM GMT+5:30
Available	dss-workshop-env	aws	US East(N. Virginia)	Running	7.2.16	4/10/2023, 8:37:03 PM GMT+5:30
Environment Stopped	cml-adb	aws	Asia Pacific (Mumbai)	Stopped	7.2.16	4/4/2023, 5:20:56 PM GMT+5:30
Available	pko-hands-on-workshop-env	aws	Asia Pacific (Mumbai)	Running	7.2.16	3/28/2023, 12:24:09 PM GMT+5:30
Environment Stopped	pse-isv-env	aws	US West (Oregon)	Stopped	7.2.16	3/23/2023, 1:22:31 AM GMT+5:30
Available	meta-workshop	aws	EU (London)	Running	7.2.16	2/24/2023, 10:30:31 PM GMT+5:30
Environment Stopped	vrayker-cdp2	aws	Asia Pacific (Sydney)	Stopped	7.2.15	10/13/2022, 5:13:00 PM GMT+5:30
Environment Stopped	pse-workshop	aws	US East (Ohio)	Stopped	7.2.16	9/16/2022, 12:32:53 AM GMT+5:30

- Click on the Data Hub associated with Streams Messaging Manager (kafka-smm-cluster)



Status	Name	Nodes	Scale	Quick Links
Running	pko-workshop-dl	2 / 0 / 0	Light Duty	Atlas Ranger Data Catalog
Running	N/A		GRN	crn:cdp:datalake:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:environment:e11adbfe-231d-4471-866a-2fb6f3352220

Data Hubs	Data Lake	FreeIPA	Cluster Definitions	Summary
3 Data Hubs				

Status	Name	Data Hub Type	Version	Node Count	Created
Running	ssb-analytics-cluster	7.2.16 - Streaming Analytics Light Duty with Apache Flink	CDH 7.2.16	6	04/20/23, 11:33 AM GMT+5:30
Running	nifi-flow-mgmt-cluster	7.2.16 - Flow Management Light Duty with Apache NiFi, Apache NiFi Registry	CDH 7.2.16	4	04/20/23, 11:21 AM GMT+5:30
Running	kafka-smm-cluster	7.2.16 - Streams Messaging Light Duty: Apache Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control	CDH 7.2.16	4	04/20/23, 11:33 AM GMT+5:30

Step 2 : Go to the Streams Messaging Interface

Data Hubs / kafka-smm-cluster / Event History

The screenshot shows the Kafka Streams Messaging Manager (SMM) interface for a cluster named 'kafka-smm-cluster'. Key details include:

- Cluster Status:** Running (4 nodes, 0 errors, 0 warnings), created at 04/20/23, 11:33 AM GMT+5:30.
- Cluster Template:** 7.2.16 - Streams Messaging Light Duty: Apache Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control.
- Environment Details:** NAME: pko-hands-on-workshop-env, DATA LAKE: pko-workshop-dl, CREDENTIAL: pko-hands-on-workshop-cred, REGION: ap-south-1, AVAILABILITY ZONE: N/A.
- Services:** CM-UI, Schema Registry, Streams Messaging Manager (highlighted with a red box), Token Integration.
- Cloudera Manager Info:** CM URL: https://kafka-smm-cluster-gateway.pko-hand.dp5i-5vkq.cloudera.site/kafka-smm-cluster/cdp-proxy/cm/home/, CM VERSION: 7.9.0, RUNTIME VERSION: 7.2.16-1.cdh7.2.16.p2.38683602, LOGS: Command logs, Service logs.
- Event History:** Autoscale, Endpoints (5), Tags (4), Nodes, Network, Load Balancers, Telemetry, Repository Details, Image Details, Recipes (0), Cloud Storage, Database, Upgrade.
- Events:** Events, Show All, Autoscale, Cluster, DOWNLOAD.

Step 3 : Select Brokers from the left tab

The screenshot shows the Kafka Streams Messaging Manager Overview page with the 'Brokers' tab selected. The interface displays the following metrics:

NAME	DATA IN	DATA OUT	MESSAGES IN	CONSUMER GROUPS	CURRENT LOG SIZE
__consumer_offsets	33 KB	33 KB	276	0	878 KB
__CruiseControlMetrics	842 KB	842 KB	76k	0	9 MB
__KafkaCruiseControlModelTrainingSamples	30 KB	0B	90	0	221 KB
__KafkaCruiseControlPartitionMetricSamples	169 KB	0B	8.8k	0	939 KB
__smm_alert_notifications	0B	0B	0	0	0B
__smm_consumer_metrics	0B	0B	0	1	0B
__smm_producer_metrics	6 KB	6 KB	57	1	104 KB

Other tabs visible include Producers (14), Topics (33), and Consumer Groups (3).

Step 4 : Save the broker list

Total Bytes In	Total Bytes Out	Produced Per Sec	Fetched Per Sec	Active Controllers	Unclean Elections	Request Pool Usage	Name	Throughput	Messages In	Partitions	Replicas	Log Size	Remaining Storage
680 MB	2 MB	307	4,928	1	0	0.00%	1546335432	682 MB	2.3m	94	261	2 GB	982 GB
							1546335453	238 KB	3.4k	100	270	6 GB	978 GB
							1546326411	500 KB	6.5k	98	263	4 GB	980 GB

Example :

kafka-smm-cluster-corebroker1.pko-hand.dp5i-5vkq.cloudera.site:9093
kafka-smm-cluster-corebroker0.pko-hand.dp5i-5vkq.cloudera.site:9093
kafka-smm-cluster-corebroker2.pko-hand.dp5i-5vkq.cloudera.site:9093

5. Download Resources from GitHub

Step 1 : Access the URL shared by the instructor for GitHub

mmehra12 / HOLWorkshops Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

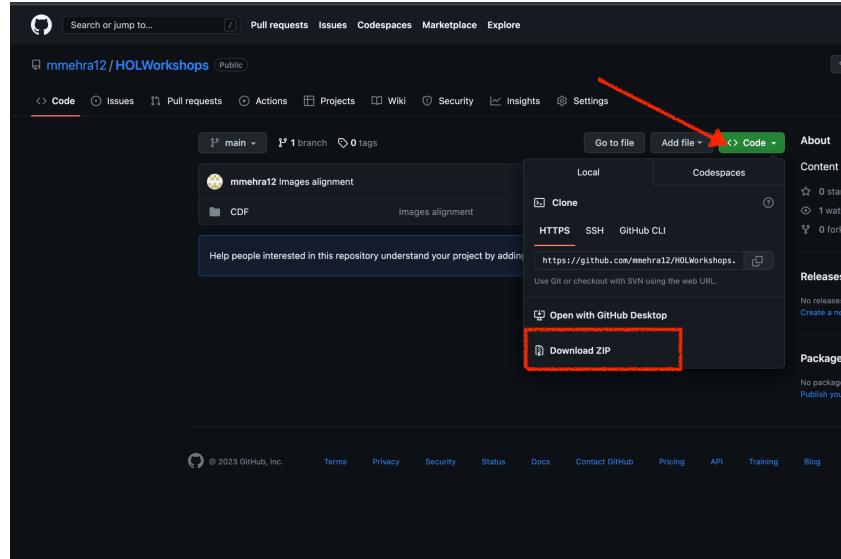
main 1 branch 0 tags Go to file Add file Code

mmehra12 Images alignment b398387 yesterday 26 commits

CDF Images alignment yesterday

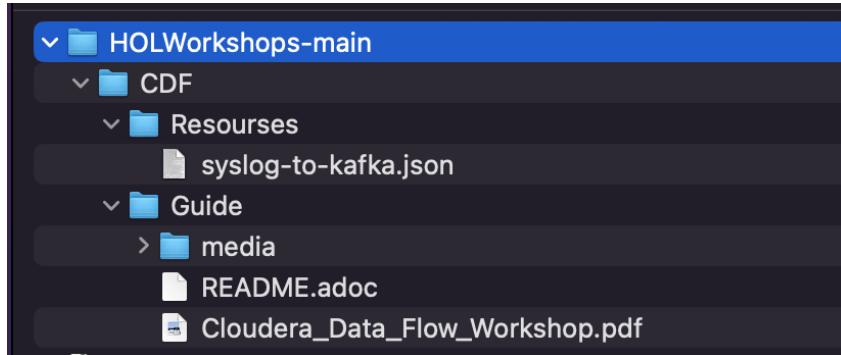
Add a README

Step 2 - Download the repo as a ZIP file



Step 3 : Uncompress the Files

Uncompress the Files and you should have the following files and folders within it



We will use this at a later point in our Labs

6. Unlock your KeyTab

To run queries on the SQL Stream Builder you need to have your KeyTab unlocked. This is mainly for authentication purposes. As the credential you are using is sometimes reused as part of other people doing the same lab it is possible that your Keytab is already unlocked. We have shared the steps for both the scenarios:

1. Unlock your Keytab if it is not unlocked already

Step 1 : Go to the SSB Data Hub

Click on Environments on the left tab and select the environment that is shared by the INSTRUCTOR

Status	Name	Cloud Provider	Region	Data Lake	CDP Runtime Version	Time Created
Available	emeworkshop-environ	aws	Asia Pacific (Mumbai)	Running	7.2.16	4/12/2023, 4:40:07 PM GMT+5:30
Available	dss-workshop-env	aws	US East(N. Virginia)	Running	7.2.16	4/10/2023, 8:37:03 PM GMT+5:30
Environment Stopped	cml-adlb	aws	Asia Pacific (Mumbai)	Stopped	7.2.16	4/4/2023, 5:20:56 PM GMT+5:30
Available	pko-hands-on-workshop-env	aws	Asia Pacific (Mumbai)	Running	7.2.16	3/28/2023, 12:24:09 PM GMT+5:30
Environment Stopped	pse-isv-env	aws	US West (Oregon)	Stopped	7.2.16	3/28/2023, 1:22:31 AM GMT+5:30
Environment Stopped	meta-workshop	aws	EU (London)	Stopped	7.2.16	2/24/2023, 10:30:31 PM GMT+5:30
Environment Stopped	vrayker-cdp2	aws	Asia Pacific (Sydney)	Stopped	7.2.15	10/13/2022, 5:13:00 PM GMT+5:30
Environment Stopped	pse-workshop	aws	US East (Ohio)	Stopped	7.2.16	9/16/2022, 12:32:53 AM GMT+5:30

Click on the DataHub associated with SQL Stream Builder (ssb-analytics-cluster)

Data Lake Details

NAME: pko-workshop-dl **NODES**: 2 (0 green, 0 yellow, 0 red) **SCALE**: Light Duty

STATUS: Running **STATUS REASON**: N/A **CRN**: cm:cdp:datalake:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:environment:e11adbfe-231d-4471-866a-2fb6f3352220

QUICK LINKS: [Atlas](#) [Ranger](#) [Data Catalog](#)

Data Hubs [Data Lake](#) [Freelipa](#) [Cluster Definitions](#) [Summary](#)

3 Data Hubs [Create Data Hub](#)

Status	Name	Data Hub Type	Version	Node Count	Created
Running	ssb-analytics-cluster	7.2.16 - Streaming Analytics Light Duty with Apache Flink	CDH 7.2.16	6	04/20/23, 11:33 AM GMT+5:30
Running	nifi-flow-mgmt-cluster	7.2.16 - Flow Management Light Duty with Apache NiFi, Apache NiFi Registry	CDH 7.2.16	4	04/20/23, 11:21 AM GMT+5:30
Running	kafka-smm-cluster	7.2.16 - Streams Messaging Light Duty: Apache Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control	CDH 7.2.16	4	04/20/23, 11:33 AM GMT+5:30

1 – 3 of 3 | < > | Items per page: 25

Step 2 : Open the SSB UI by clicking on **Streaming SQL Console**

Data Hubs / ssb-analytics-cluster / Event History

ssb-analytics-cluster [Actions](#)

cm:cdp:datalib:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:cluster:ca445db-316f-4735-96cc-4ebd0ddc5750

STATUS	NODES	CREATED AT	CLUSTER TEMPLATE	STATUS REASON
Running	6 (0 green, 0 yellow, 0 red)	04/20/23, 11:33 AM GMT+5:30	7.2.16 - Streaming Analytics Light Duty with Apache Flink	Cluster started.

Environment Details

NAME	DATA LAKE	CREDENTIAL	REGION	AVAILABILITY ZONE
pko-hands-on-workshop-env	pko-workshop-dl	pko-hands-on-workshop-cred	ap-south-1	N/A

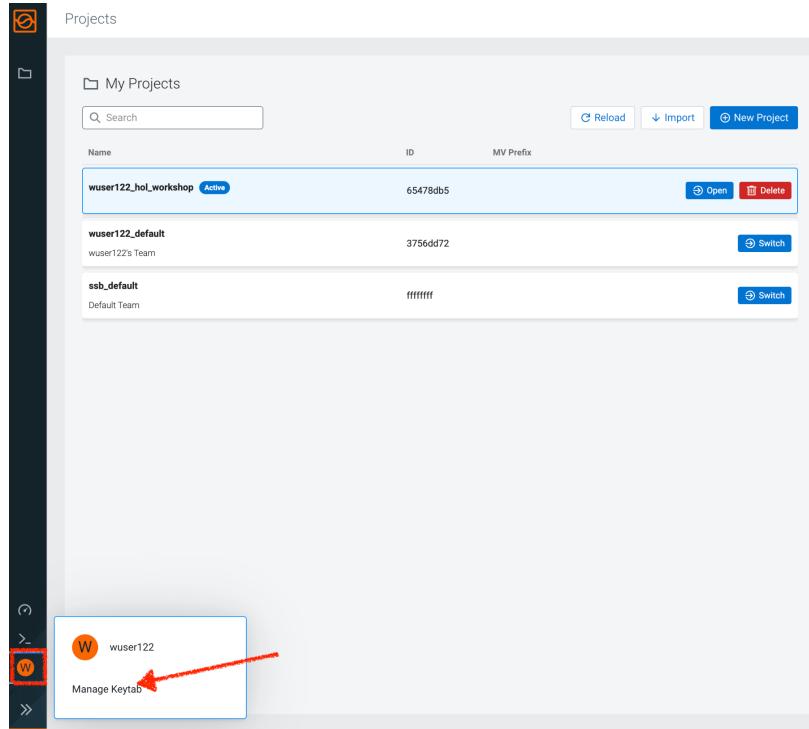
Services

CM-UI	Flink Dashboard	Job History Server	Name Node	Name Node
Queue Manager	Resource Manager	Streaming SQL Console	Token Integration	

Cloudera Manager Info

CM URL	CM VERSION	RUNTIME VERSION	LOGS
https://ssb-analytics-cluster-gateway.pko-hand.dp5i-5vkq.cloudera.site/ssb-analytics-cluster/cdp-proxy/cm/home/	7.9.0	7.2.16-1.cdh7.2.16.p2.38683602	Command logs , Service logs

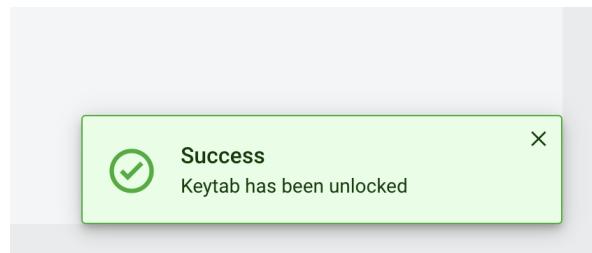
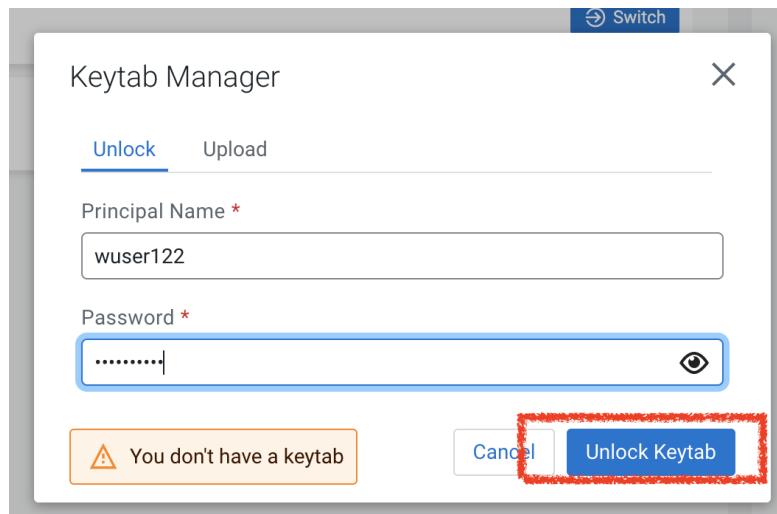
Step 3 : Click on the User name at the bottom left of the screen and select Manage Keytab



Step 4 : Enter your Workload Username (apacXY) and Password.

The screenshot shows a 'Keytab Manager' dialog box. At the top right is a 'Switch' button. The dialog has tabs for 'Unlock' (selected) and 'Upload'. Below are fields for 'Principal Name *' and 'Password *', both with red asterisks indicating required fields. A message at the bottom states '⚠ You don't have a keytab'. At the bottom right are 'Cancel' and 'Unlock Keytab' buttons.

Step 5 : Click on unlock KeyTab



2. Reset your KeyTab if it is already unlocked

Step 1 : Go to the SSB Data Hub

Click on Environments on the left tab and select the environment that is shared by the INSTRUCTOR

Status	Name	Cloud Provider	Region	Data Lake	CDP Runtime Version	Time Created
Available	emeaworkshop-env	aws	Asia Pacific (Mumbai)	Running	7.2.16	4/12/2023, 4:40:07 PM GMT+5:30
Available	dss-workshop-env	aws	US East(N. Virginia)	Running	7.2.16	4/10/2023, 8:37:03 PM GMT+5:30
Environment Stopped	cml-sdb	aws	Asia Pacific (Mumbai)	Stopped	7.2.16	4/4/2023, 5:20:56 PM GMT+5:30
Available	pko-hands-on-workshop-env	aws	Asia Pacific (Mumbai)	Running	7.2.16	3/28/2023, 12:24:09 PM GMT+5:30
Environment Stopped	pse-isv-env	aws	US West (Oregon)	Stopped	7.2.16	3/23/2023, 1:22:31 AM GMT+5:30
Environment Stopped	meta-workshop	aws	EU (London)	Stopped	7.2.16	2/24/2023, 10:30:31 PM GMT+5:30
Environment Stopped	vrayker-cdp2	aws	Asia Pacific (Sydney)	Stopped	7.2.15	10/13/2022, 5:13:00 PM GMT+5:30
Environment Stopped	pse-workshop	aws	US East (Ohio)	Stopped	7.2.16	9/16/2022, 12:32:53 AM GMT+5:30

Click on the DataHub associated with SQL Stream Builder (ssb-analytics-cluster)

Status	Name	NODES	SCALE	QUICK LINKS
Running	pko-workshop-dl	2 0 0	Light Duty	Atlas Ranger Data Catalog
Running	ssb-analytics-cluster	2 0 0	Light Duty	Atlas Ranger Data Catalog
Running	nifi-flow-mgmt-cluster	2 0 0	Light Duty	Atlas Ranger Data Catalog
Running	kafka-smm-cluster	2 0 0	Light Duty	Atlas Ranger Data Catalog

Step 2 : Open the SSB UI by clicking on Streaming SQL Console

Data Hubs / ssb-analytics-cluster / Event History

ssb-analytics-cluster

cm.cdp.datahub.us-west-1.d1a4553c-a799-432d-8e54-372cc2ab95f2:cluster.ca4445db-316f-4735-96cc-4ebd0ddc5750

STATUS	NODES	CREATED AT	CLUSTER TEMPLATE	STATUS REASON
Running	6 0 0	04/20/23, 11:33 AM GMT+5:30	7.2.16 - Streaming Analytics Light Duty with Apache Flink	Cluster started.

aws Environment Details

NAME pko-hands-on-workshop-env	DATA LAKE pko-workshop-dl	CREDENTIAL pko-hands-on-workshop-cred	REGION ap-south-1	AVAILABILITY ZONE N/A
-----------------------------------	------------------------------	--	----------------------	--------------------------

Services

CM CM-UI	Flink Dashboard	Job History Server	Name Node	Name Node
Queue Manager	Resource Manager	Streaming SQL Console	Token Integration	

Cloudera Manager Info

CM URL https://ssb-analytics-cluster-gateway.pko-hand.dp5i-5vkq.cloudera.site/ssb-analytics-cluster/cdp-proxy/cm/home/	CM VERSION 7.9.0	RUNTIME VERSION 7.2.16-1.cdh7.2.16.p2.38683602	LOGS Command logs , Service logs
---	---------------------	---	---

Step 3 : Click on the User name at the bottom left of the screen and select Manage Keytab

Projects

My Projects

Name	ID	MV Prefix
wuser122_hol_workshop	65478b05	
wuser122_default	3756dd72	
ssb_default	fffffff	

wuser122
Manage Keytab

Keytab Manager

Lock Upload

Principal Name *

You have a keytab

Cancel Lock Keytab

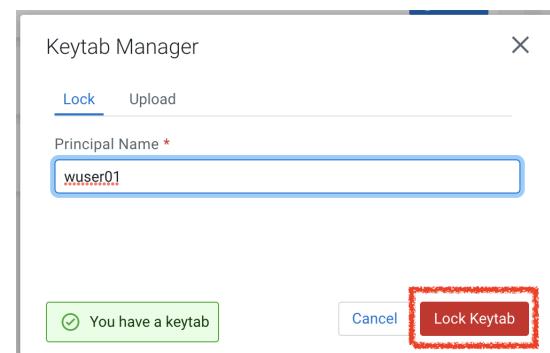
If you get the following dialog box it means that your Keytab is already unlocked. **But it would be necessary to reset here by locking it and unlocking it again using your newly set workload password**

Step 3 : Enter your Principal Name which is the same as your workload username

Example : apacXY

Click on Lock KeyTab

You can now continue from the STEP 3 in
the "[Unlock your KeyTab if not unlocked already](#)"
section above



Lab 1 : Create a Flow using the Flow Designer

1. Overview

Creating a data flow for CDF-PC is the same process as creating any data flow within Nifi with 3 very important steps:

- The data flow that would be used for CDF-PC must be self contained within a process group
- Data flows for CDF-PC must use parameters for any property on a processor that is modifiable, e.g. user names, Kafka topics, etc.
- All queues need to have meaningful names (instead of Success, Fail, and Retry). These names will be used to define Key Performance Indicators in CDF-PC.

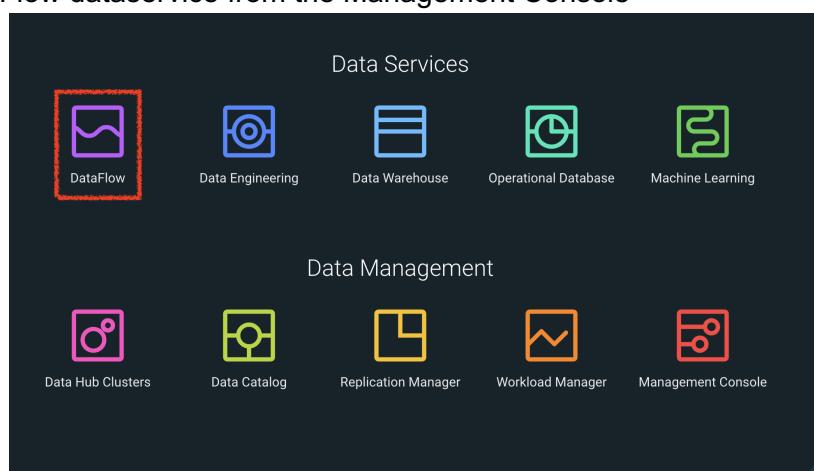
The following is a step by step guide in building a data flow for use within CDF-PC.

2. Building the Data Flow

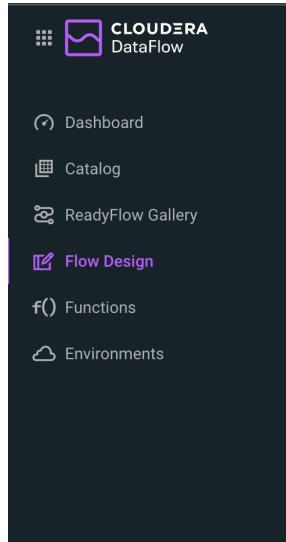
2.1. Create the canvas to design your flow

Step 1: Access the DataFlow Data Service

Access the DataFlow dataservice from the Management Console



Step 2: Go to the Flow Design



Step 3: Create a new Draft

(This will be the main process group of your flow)

Flow Design

All Drafts

Search for a draft by name

Create Draft

REFRESHED: 7 seconds ago

A screenshot of the Cloudera DataFlow 'Flow Design' interface. It shows a list of 'All Drafts' with a search bar. A prominent blue button labeled 'Create Draft' is visible, with a red box drawn around it to indicate it as the next step.

Step 4: Select the appropriate environment

Select the appropriate environment as part of the workspace and give your flow a name and click on **CREATE**

Workspace Name : *The name of the environment will be shared by the INSTRUCTOR*

Draft Name : {user_id}_datadump_flow
Example : apacXY_datadump_flow

Create New Draft

Select the target workspace

Workspace [?](#)

aws pko-hands-on-workshop-env 15% (3 of 20)

Draft Name

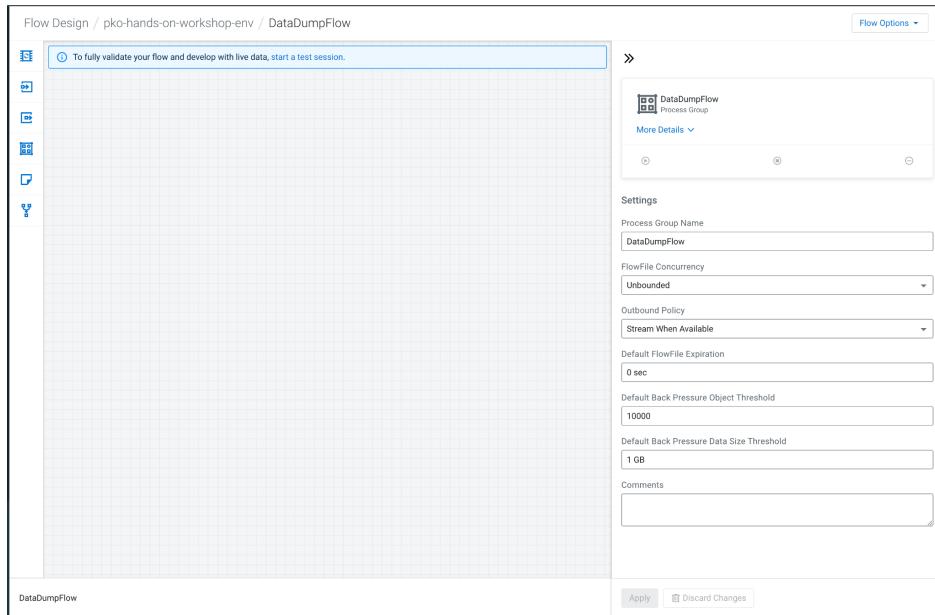
wuser00_datadump_flow

Draft name is valid

Cancel Create

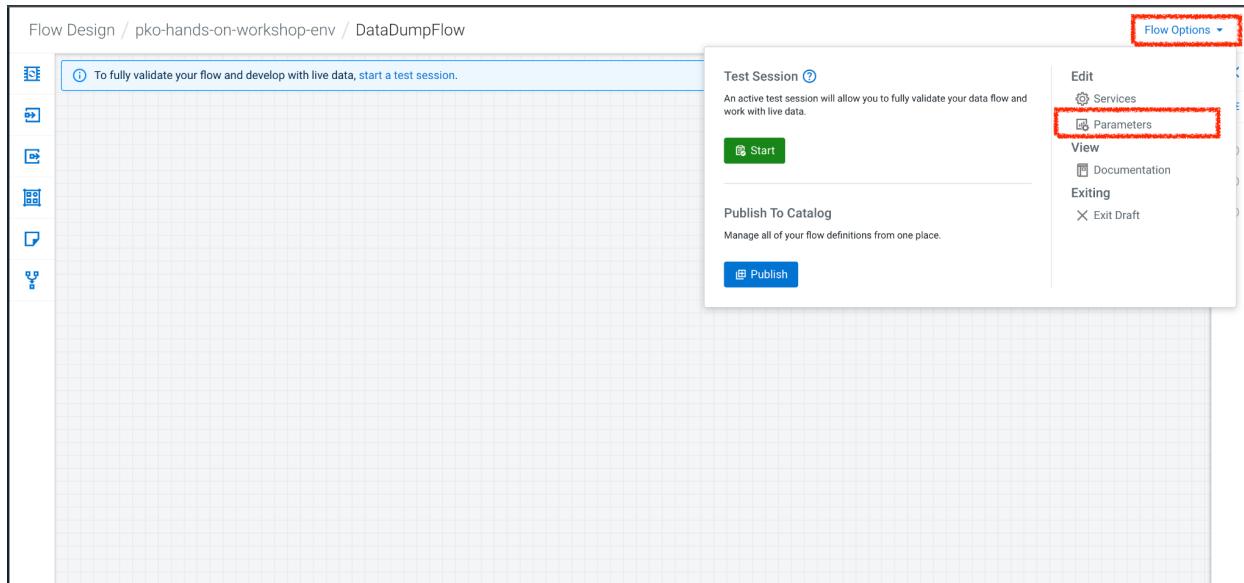
A screenshot of the 'Create New Draft' dialog box. It asks for the 'target workspace' and shows 'aws pko-hands-on-workshop-env' selected. Below, the 'Draft Name' field contains 'wuser00_datadump_flow', which is highlighted with a blue border. A green checkmark indicates the name is valid. At the bottom are 'Cancel' and 'Create' buttons.

On successful creation of the Draft, you should now be redirected to the canvas on which you can design your flow



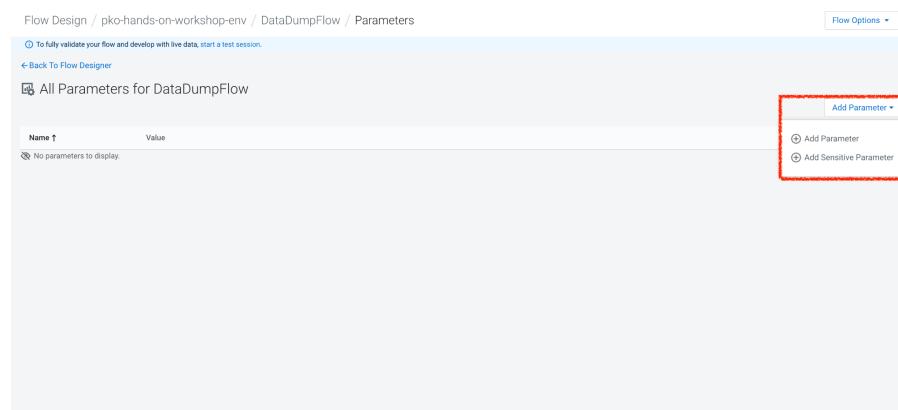
2.2. Adding new parameters

Step 1: Click on the **FLOW OPTIONS** on the top right corner of your canvas and then select **PARAMETERS**



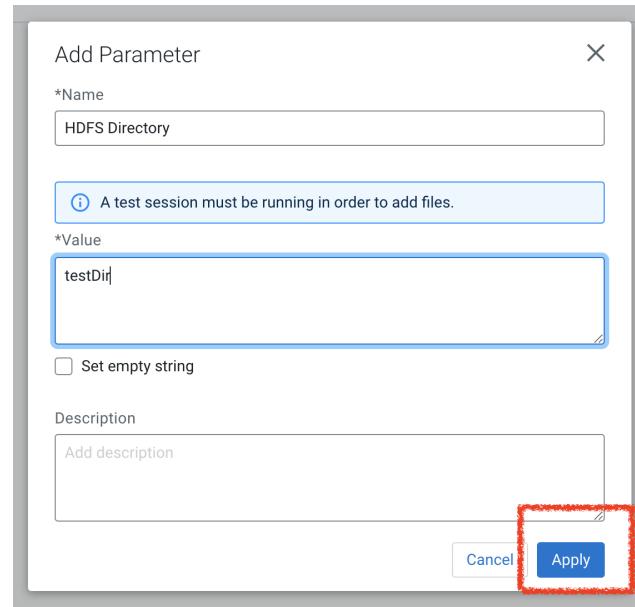
Step 2: Configure Parameters

The next step is to configure what is called a parameter. These parameters are reused within the flow multiple times and will also be configurable at the time of deployment. Click on **ADD PARAMETER** to add non sensitive values, for any sensitive parameter please select **ADD SENSITIVE PARAMETER**.

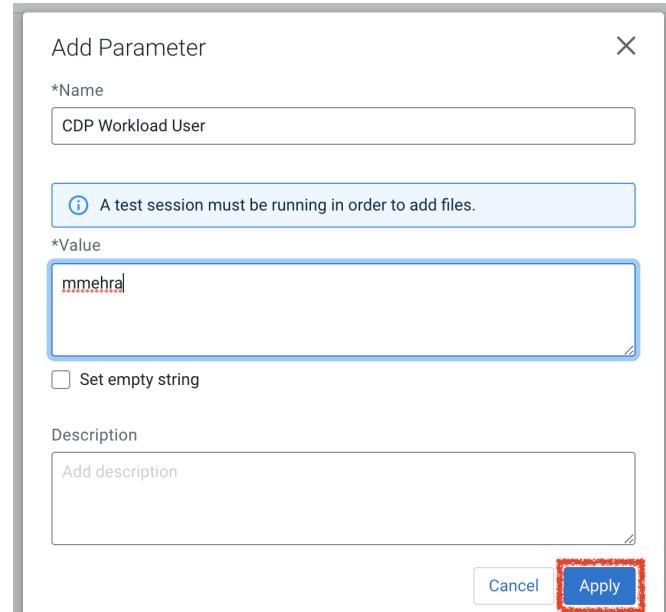


We need to add the following parameters.

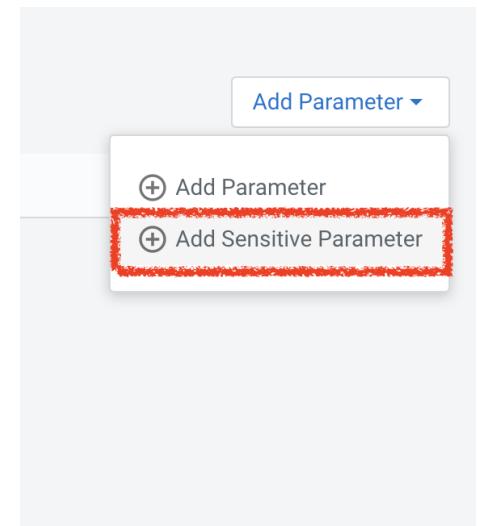
- S3 Directory
 - Selection under Add Parameter : **Add Parameter**
 - Name : S3 Directory
 - Value : LabData or TestDir



- CDP Workload User
 - Selection under Add Parameter : **Add Parameter**
 - Name : CDP Workload User
 - Value : <The username assigned to you>
 - EXAMPLE : apac01
 - **IMPORTANT:** do not add the domain '@workload.com'



- CDP Workload User Password - [Sensitive Field]
 - Selection under Add Parameter : **Add Sensitive Parameter**
 - Name : CDP Workload User Password
 - Value : <Workload Password set by yourself in Lab 0>
 - EXAMPLE : ApacLabs@23



Add Sensitive Parameter X

*Name
CDP Workload User Password

i A test session must be running in order to add files.

*Value
.....
 Set empty string

Description
Add description

Cancel Apply

← Back To Flow Designer

All Parameters for hostmm_datadump_flow

Name ↑	Value	Changed
CDP Workload User	wuser00	Modified >
CDP Workload User Password	Sensitive value set	Modified >
S3 Directory	LabData	Modified >

Apply Changes Discard Changes

Click **APPLY CHANGES**

Now go back to the Flow Designer. Click '*Back to Flow Designer*'

To fully validate your flow and develop with live data, start a test session.

< Back To Flow Designer

All Parameters for wuser122_datadump_flow

Name ↑	Value	Changed
CDP Workload User	wuser122	
CDP Workload User Password	Sensitive value set	
S3 Directory	wrkshop_data	

Add Parameter ▾

Now that we have created these parameters, we can easily search and reuse them within our dataflow. This is especially useful for **CDP Workload User** and **CDP Workload User Password**.

NOTE ONLY:

To search for existing parameters:

1. Open a processor's configuration and proceed to the properties tab.
2. Enter: #{{
3. Hit 'control+spacebar'

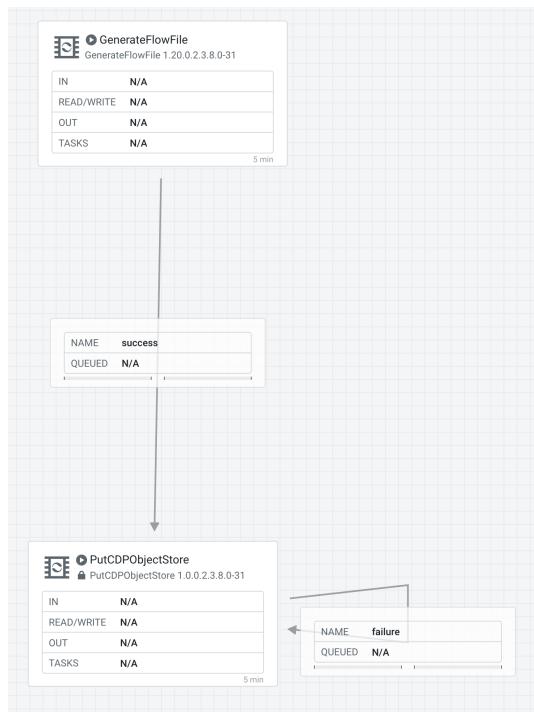
This will bring up a list of existing parameters that are not tagged as sensitive.

2.3. Create the Flow

Let's go back to the canvas to start designing our flow. This flow will contain 2 Processors:

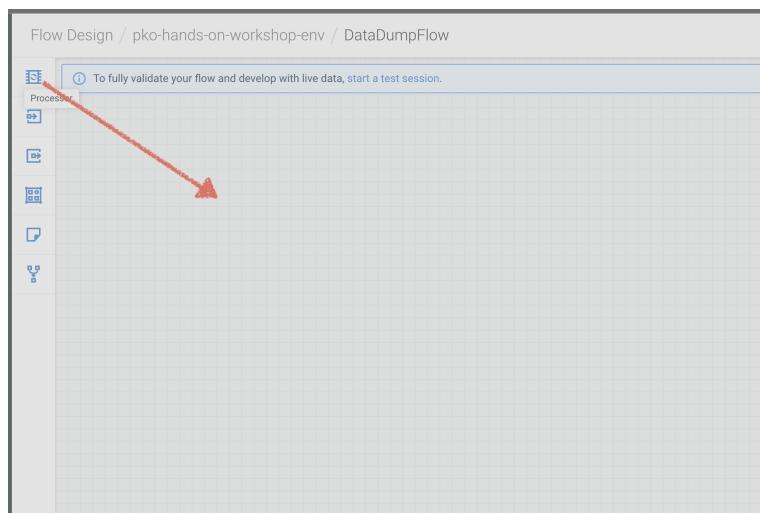
- **GenerateFlowFile** - Generates random data
- **PutCDPObjectStore** - Loads data into HDFS(S3)

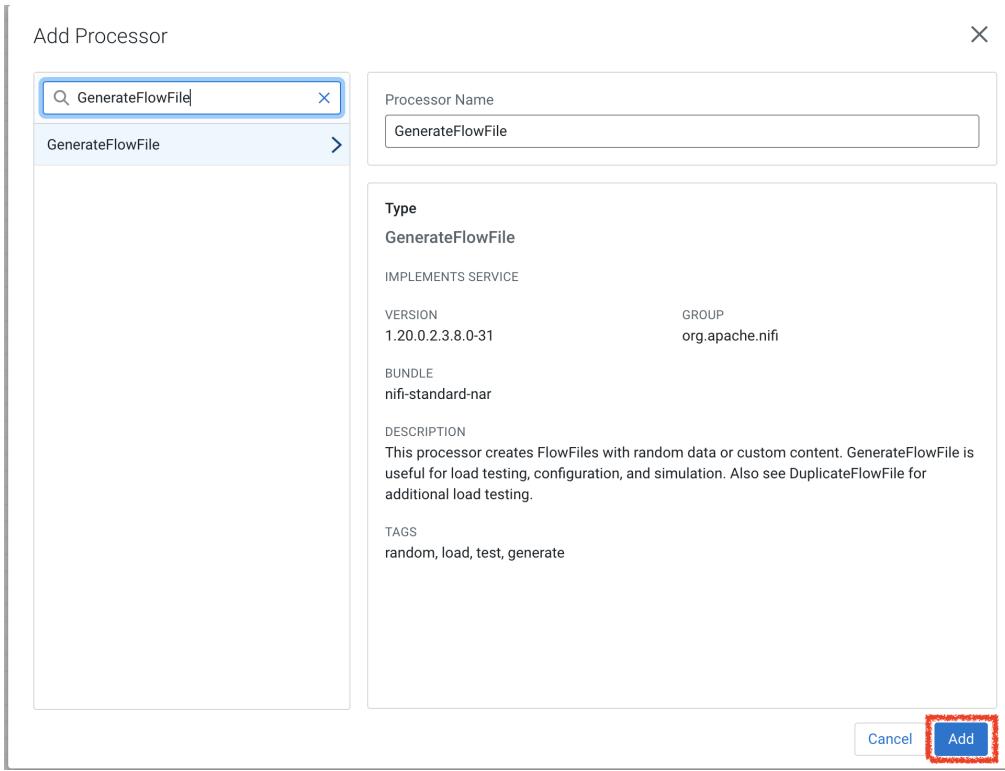
Our final flow will look like this:



Step 1: Add **GenerateFlowFile** processor

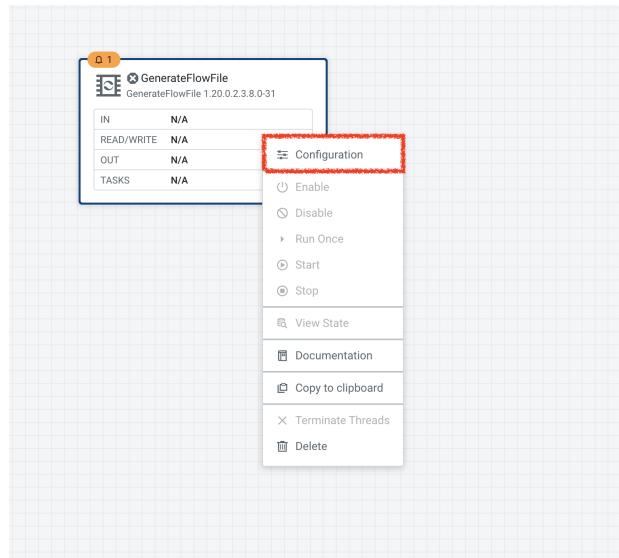
Pull the Processor onto the canvas and select **GenerateFlowFile Processor** and click on **ADD**.





Step 2: Configure GenerateFlowFile processor

The GenerateFlowFile Processor will now be on your canvas and you can configure it in the following way by right clicking and selecting **Configuration**.



Configure the processor in the following way

Property	Value
Processor Name	DataGenerator
Scheduling Strategy(default)	Timer Driven
Run Duration(default)	0 ms
Run Schedule	30 sec
Execution(default)	All Nodes
Custom Text	<26>1 2021-09-21T21:32:43.967Z host1.example.com application4 3064 ID42 [exampleSDID@873 iut="4" eventSource="application" eventId="58"] application4 has stopped unexpectedly

This represents a syslog out in RFC5424 format. Subsequent portions of this workshop will leverage this same syslog format.

»

 **GenerateFlowFile**
 GenerateFlowFile 1.20.0.2.3.8.0-31

[More Details](#)

⌚ ⌚ ⌚ ⌚ ⌚ ⌚ ⌚ ⌚ ⌚ ⌚ ⌚ ⌚

Settings

*Processor Name

*Penalty Duration ? *Yield Duration ?

*Bulletin Level ?

Comments

Scheduling

*Scheduling Strategy ? *Concurrent Tasks ?

*Run Duration ? *Run Schedule ?

*Execution ?

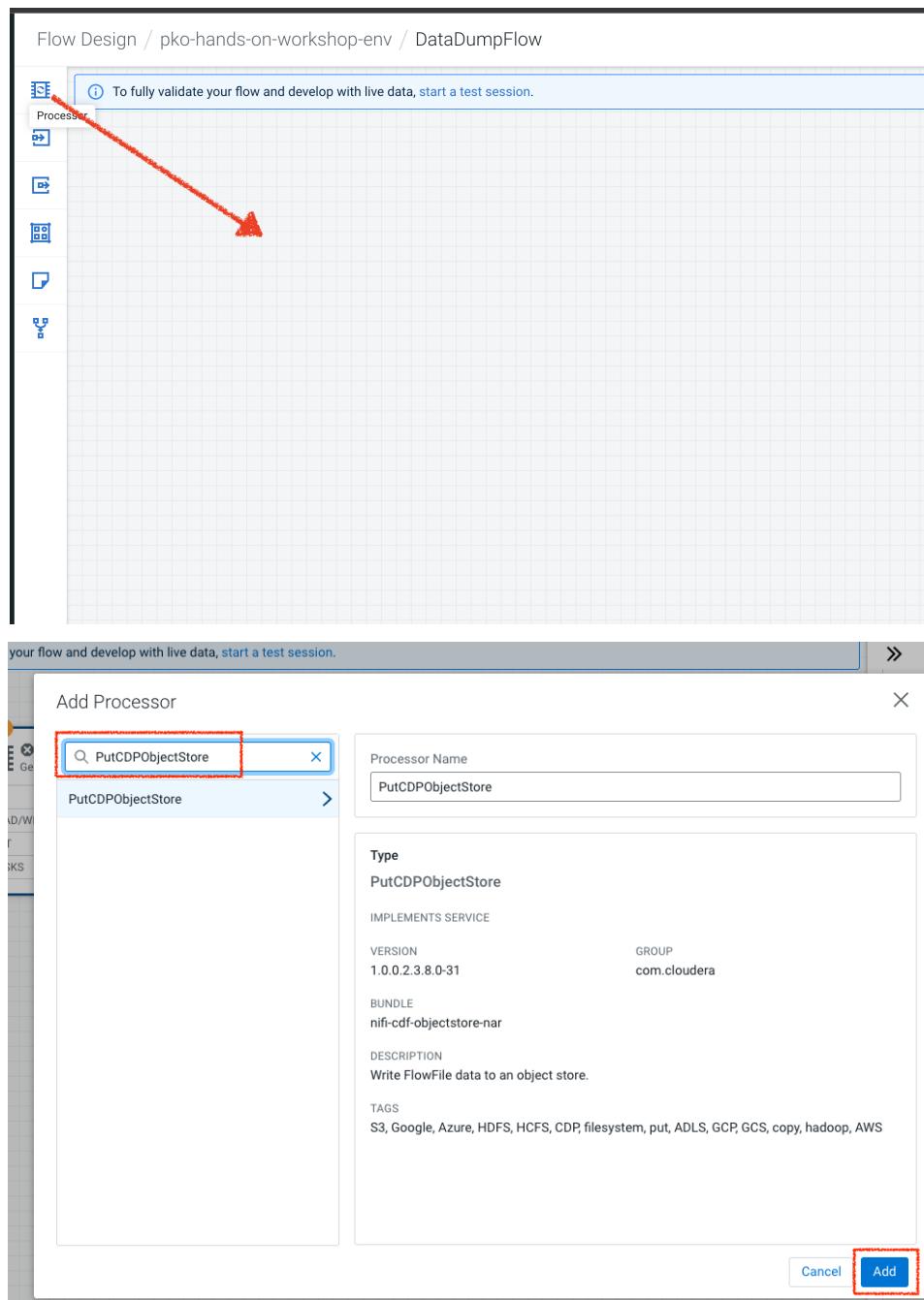
Properties ⊕ Add Property

Property	Value	⋮
File Size ?	0B	⋮
Batch Size ?	1	⋮
Data Format ?	Text	⋮
Unique FlowFiles ?	false	⋮
Custom Text ?	<26>1 2021-09-21T21:32:43.967Z ...	⋮
Character Set ?	UTF-8	⋮
Mime Type ?	No value set	⋮

Click on **APPLY**.

Step 3: Add PutCDPObjectStore processor

Pull a new Processor onto the canvas and select **PutCDPObjectStore** Processor and click on **ADD**.



Step 4: Configure PutCDPObjectStore processor

The PutCDPObjectStore Processor needs to be configured as follows:

Property	Value
Processor Name	Move2S3
Scheduling Strategy(default)	Timer Driven
Run Duration(default)	0 ms
Run Schedule(default)	0 sec
Execution(default)	All Nodes
Directory	#{S3 Directory}
CDP Username	#{CDP Workload User}
CDP Password	#{CDP Workload User Password}
Auto Terminate Relationships:	Check the “Terminate” box under “success”

More Details ▾



Settings

*Processor Name

Move2S3

*Penalty Duration ②

30 sec

*Yield Duration ②

1 sec

*Bulletin Level ②

WARN

Comments

Scheduling

*Scheduling Strategy ②

Timer Driven

*Concurrent Tasks ②

1

success ②

Terminate

Retry

*Run Duration ②

0ms

*Run Schedule ②

0 sec

failure ②

Terminate

Retry

*Execution ②

All Nodes

Retry logic specified below will apply to all relationships for this processor that are set to retry.

*Number of Retry Attempts ②

10

You can choose to automatically terminate and/or retry FlowFiles sent to a given relationship if it is not defined elsewhere. If both terminate and retry are selected, retry logic will occur first, followed by termination.

success ②

Terminate Retry

failure ②

Terminate Retry

Retry logic specified below will apply to all relationships for this processor that are set to retry.

*Number of Retry Attempts ②

10

*Retry Back Off Policy ②

Penalize Yield

*Retry Maximum Back Off Period ②

10 mins

Apply

 Discard Changes

Properties

 Add Property

Property	Value	⋮
Storage Location ②	No value set	⋮
Directory ②	↪ #{HDFS Directory}	⋮
Conflict Resolution Strategy ②	fail	⋮
Kerberos Credentials Service ②	No value set	⋮
CDP Username ②	↪ #{CDP Workload User}	⋮
CDP Password ②	↪ #{CDP Workload User Password}	⋮
Writing Strategy ②	Simple write	⋮
cdp.configuration.resources ②	#(CDPEnvironment)	⋮

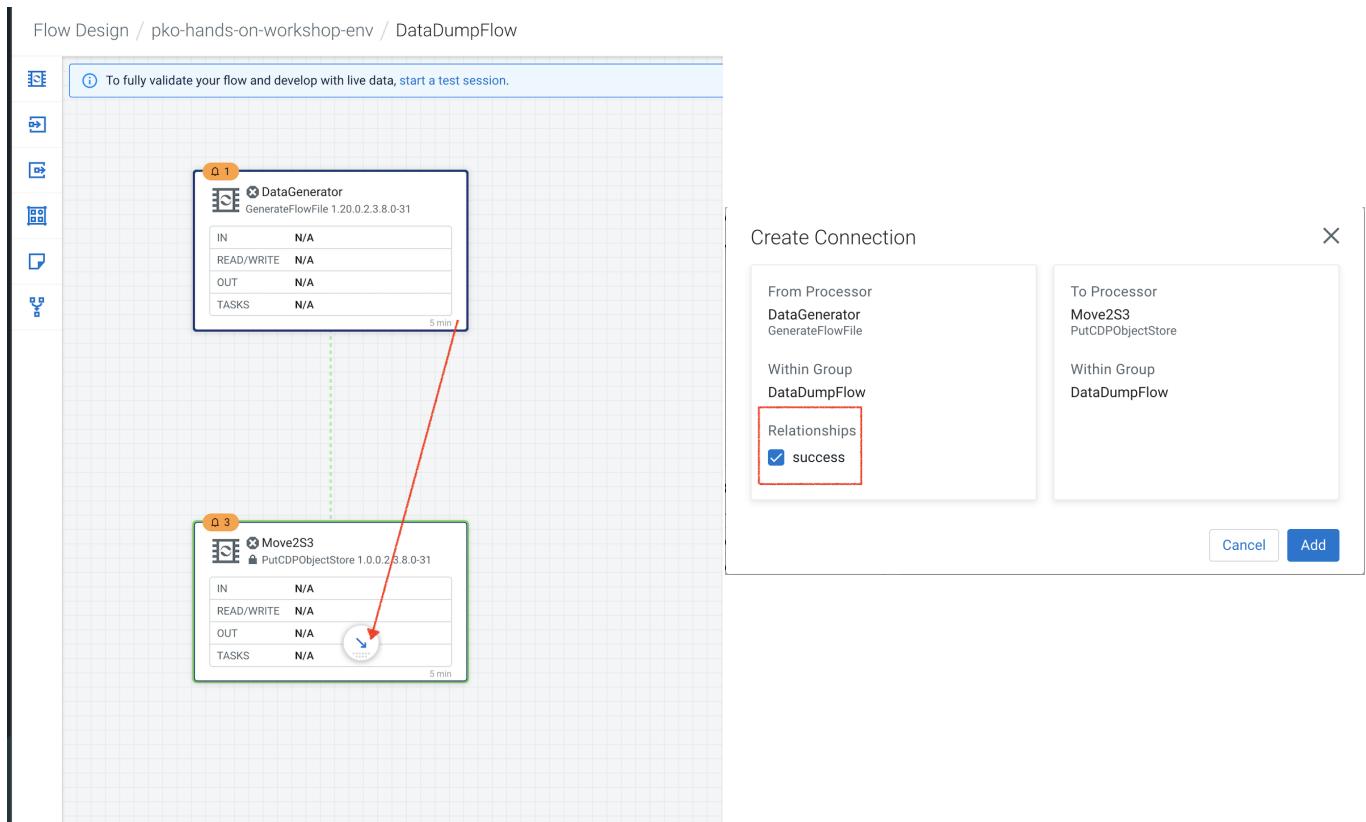
Relationships

You can choose to automatically terminate and/or retry FlowFiles sent to a given relationship if it is not defined elsewhere. If both terminate and retry are selected, retry logic will occur first, followed by termination.

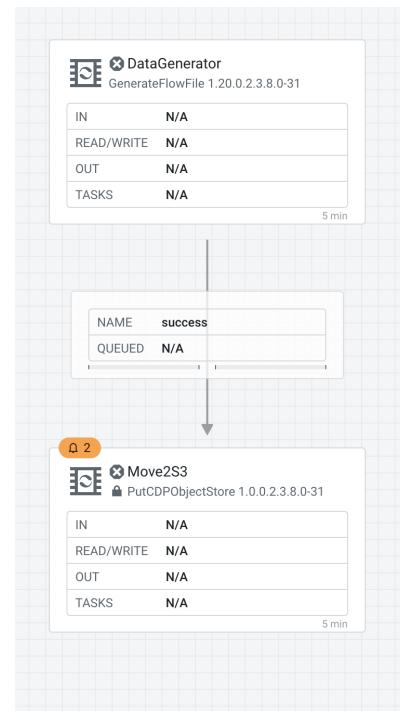
Click APPLY

Step 5: Create connection between processors

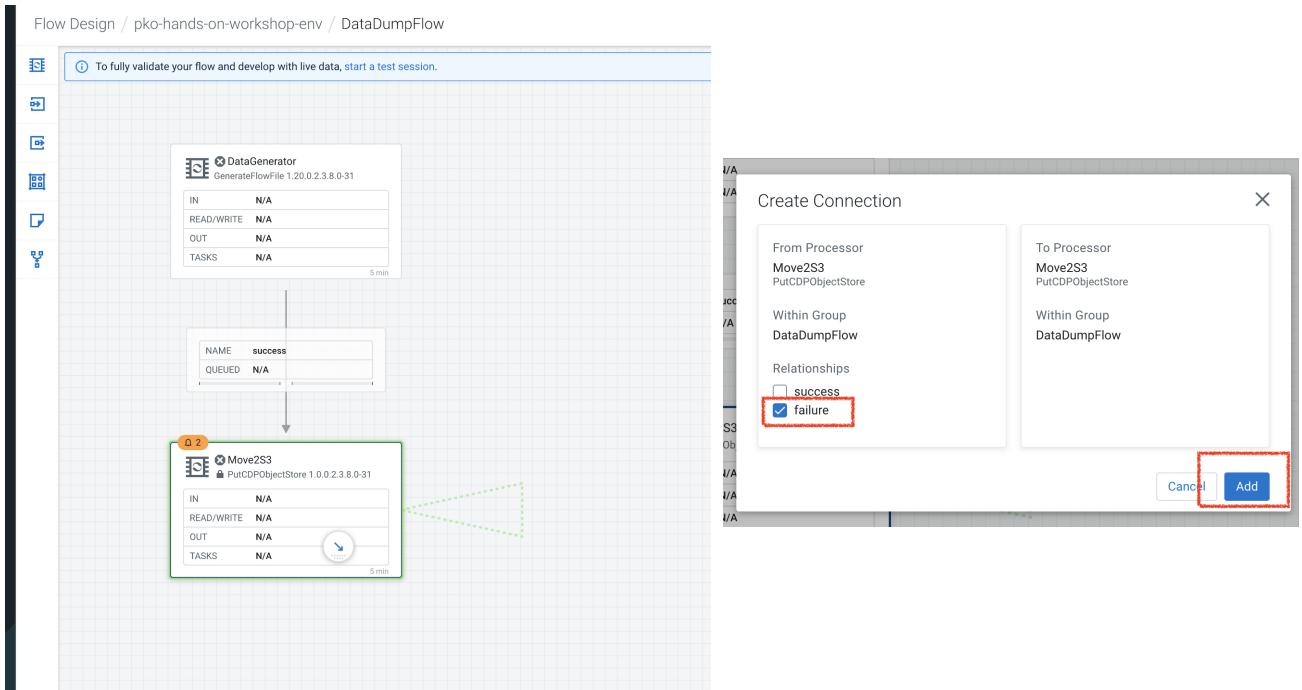
Connect the two processors by dragging the arrow from **DataGenerator** processor to the **Move2S3** processor and select on **SUCCESS** relation and click **ADD**

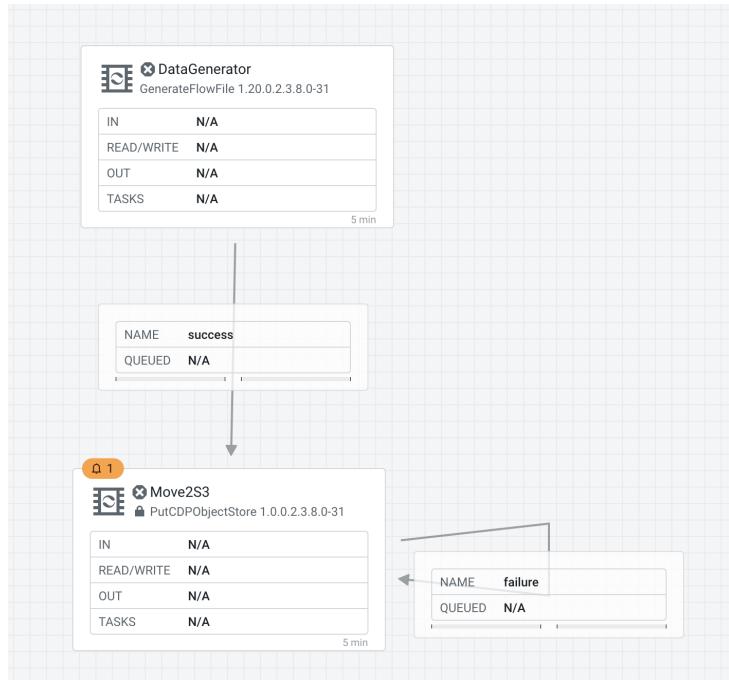


Your flow will now look something like this:



The Move2S3 processor does not know what to do in case of a failure, let's add a retry queue to it. This can be done by dragging the arrow on the processor outwards then back to itself, as below:



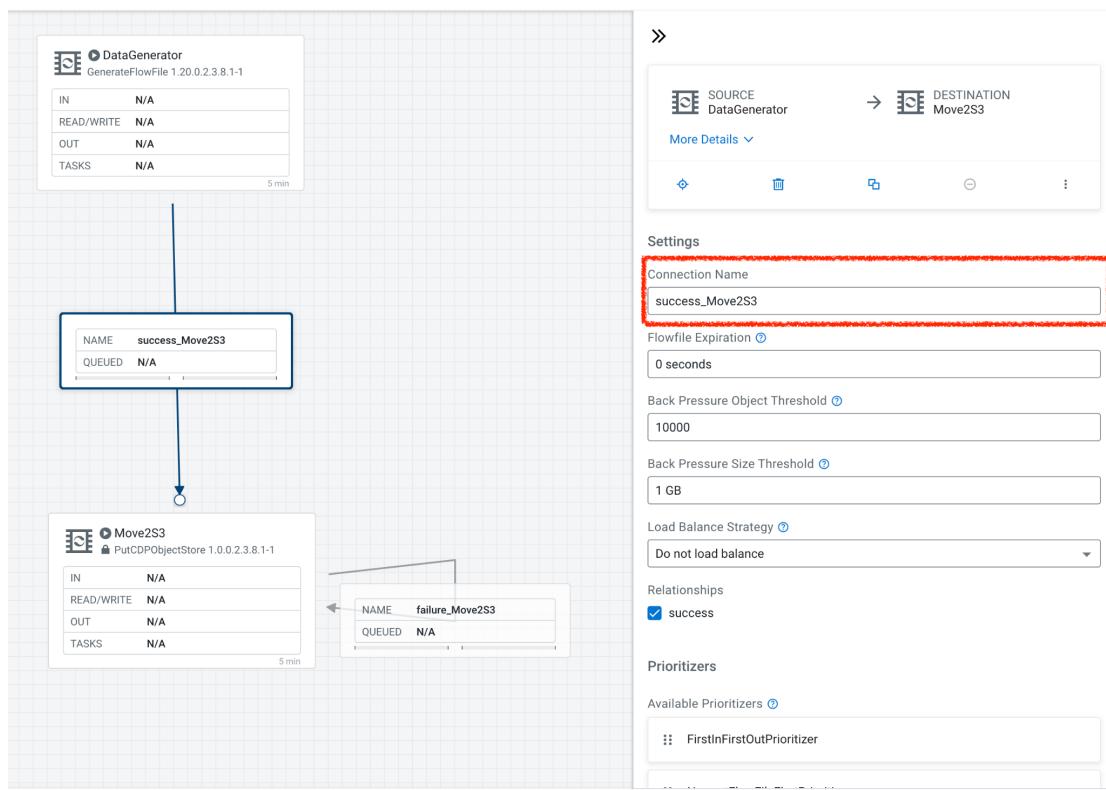


2.4. Naming the queues

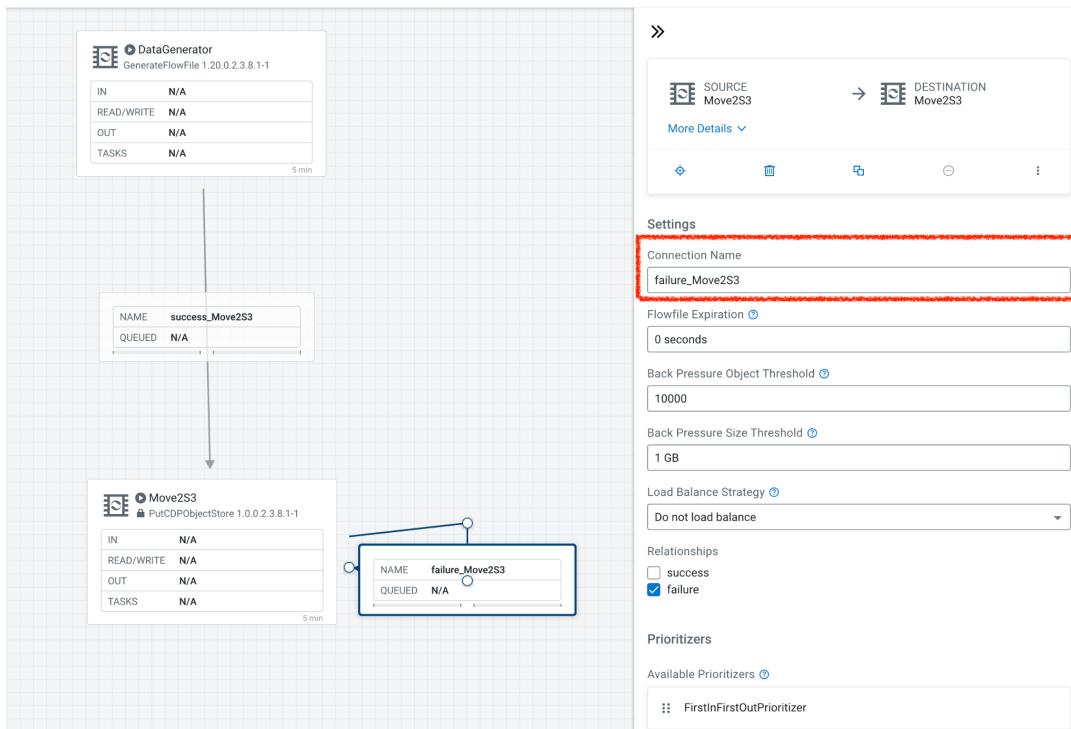
Providing unique names to all queues is very important as they are used to define Key Performance Indicators upon which CDF-PC will auto-scale.

To name a queue, double-click the queue and give it a unique name. A best practice here is to start the existing queue name (i.e. success, failure, retry, etc...) and add the source and destination processor information.

For example, the success queue between **DataGenerator** and **Move2S3** is named **success_Move2S3**.



The failure queue for Move2S3 is named **failure_Move2S3**.

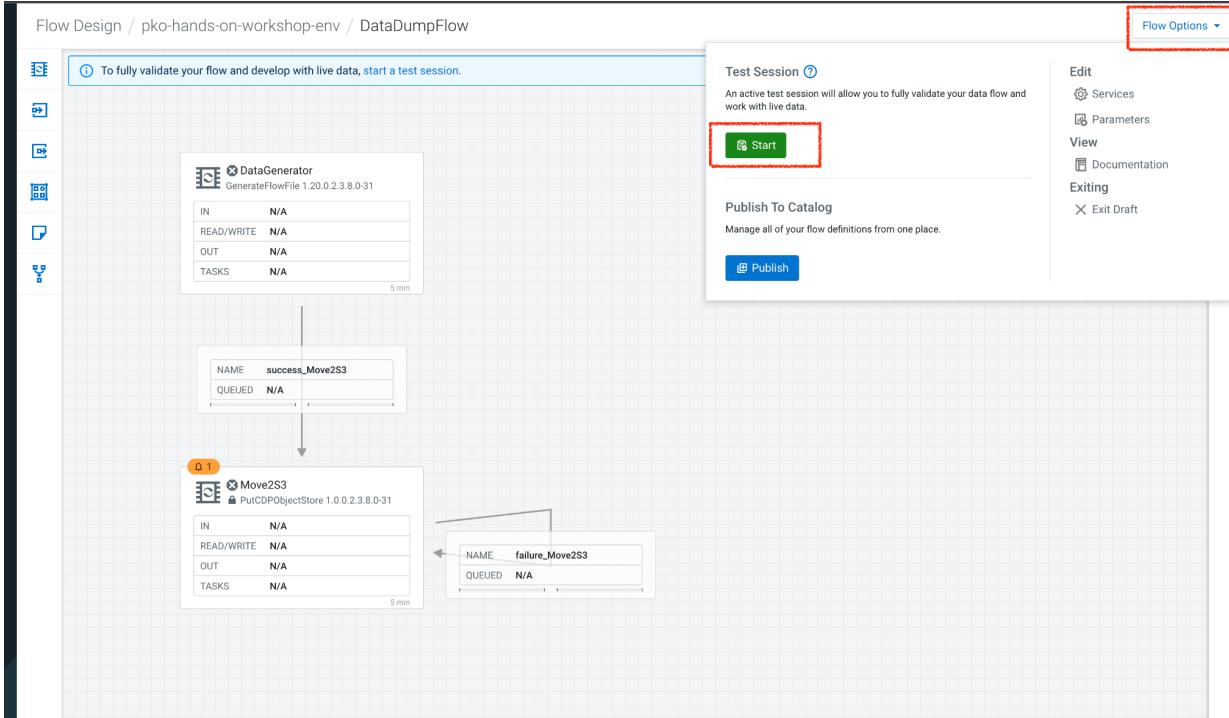


3. Testing the Data Flow

Step 1: Start test session

To test your flow we need to first start the test session

Click on **FLOW OPTIONS** and then select **START** on TEST SESSION



In the next window, click **START SESSION**

Flow Design / pko-hands-on-workshop-env / DataDumpFlow / Test Session

The activation should take about a couple of minutes. While this happens you will see this at the top right corner of your screen

⌚ Initializing Test Session...

Flow Options ▾

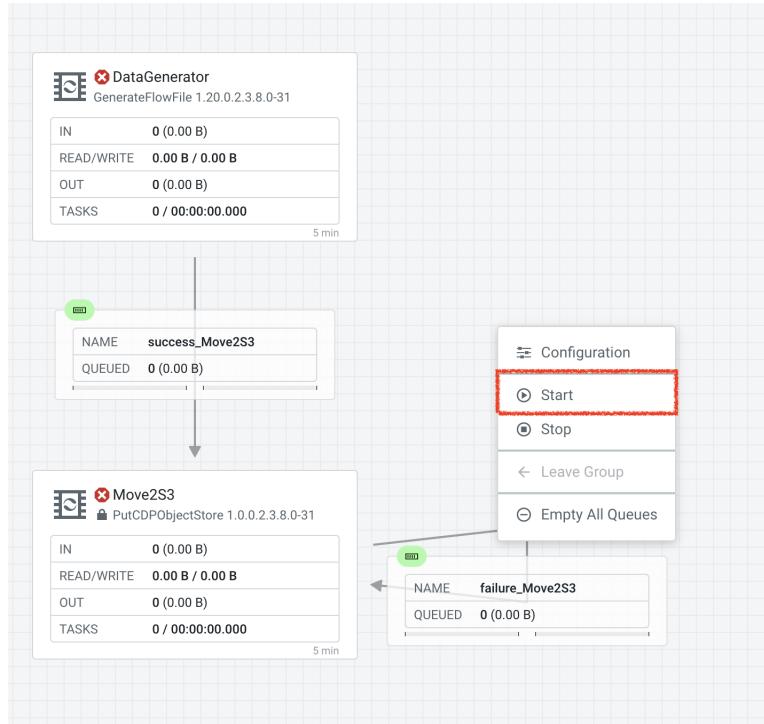
Once the Test Session is ready you will see the following message on the top right corner of your screen.

Active Test Session

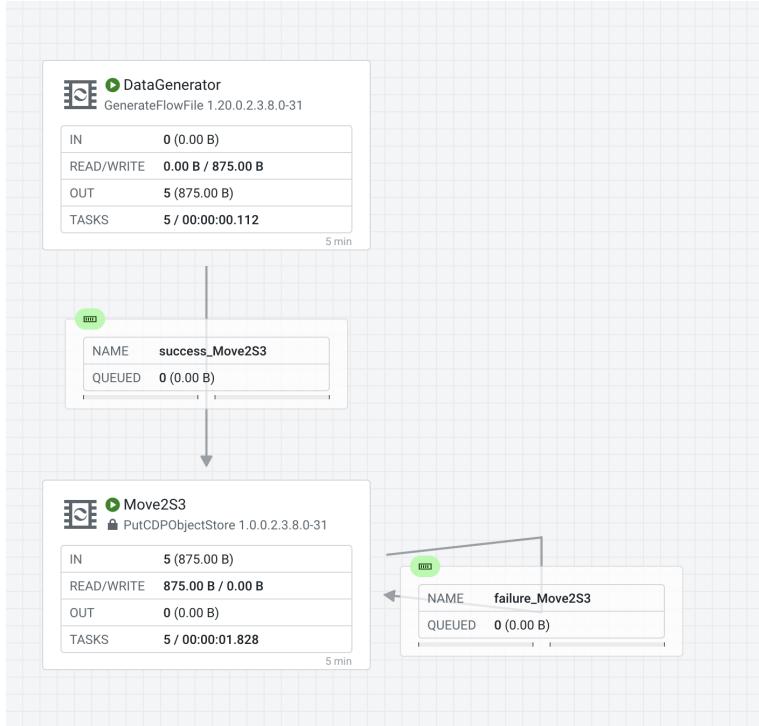
Flow Options ▾

Step 2: Run the flow

Right click on the empty part of the canvas and select START.



Both the processors should now be in the START state.



You will now see files coming into the folder which was specified as the Directory on the S3 bucket which is the Base data store for this environment.

Name ↑	Value
CDP Workload User	mmehra
CDP Workload User Password	☒ Sensitive value set
CDPEnvironment	hive-site.xml, core-site.xml, ssl-client.xml
Default SSL Context Keystore	/home/nifi/additional/secret/ssl_keystore/ssl-keystore.jks
Default SSL Context Keystore Password	☒ Sensitive value set
Default SSL Context Keystore Type	JKS
Default SSL Context Truststore	/home/nifi/additional/secret/ssl_truststore/ssl-truststore.jks
Default SSL Context Truststore Password	☒ Sensitive value set
Default SSL Context Truststore Type	JKS
HDFS Directory	newtest

Amazon S3 > Buckets > handsonworkshop > user > mmehra / newtest/

newtest/

Copy S3 URI

Objects Properties

Objects (9)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 Inventory [\[?\]](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more \[?\]](#)

Actions Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

Name	Type	Last modified	Size	Storage class
2d08ec54-124d-4f39-8536-655589724752	-	March 29, 2023, 20:51:27 (UTC+05:30)	175.0 B	Standard
3158ab72-5344-469c-966d-76f3a5e81170	-	March 29, 2023, 20:53:27 (UTC+05:30)	175.0 B	Standard
3a530f32-fec7-4821-83b8-ca79c4e11e21	-	March 29, 2023, 20:53:57 (UTC+05:30)	175.0 B	Standard
5e6348d5-0100-4568-ac49-0c6127968aff	-	March 29, 2023, 20:52:27 (UTC+05:30)	175.0 B	Standard
69814fcf-9e21-414e-b08c-10895e7fd08b	-	March 29, 2023, 20:52:57 (UTC+05:30)	175.0 B	Standard
b8fc998a-21d1-4f8b-8072-5ad350b74135	-	March 29, 2023, 20:54:27 (UTC+05:30)	175.0 B	Standard
acc85d58-9aaa-4451-816a-7901ef6b65bf	-	March 29, 2023, 20:54:57 (UTC+05:30)	175.0 B	Standard
c8a5becb-6d4e-430c-9dfe-c477d992e30c	-	March 29, 2023, 20:51:24 (UTC+05:30)	175.0 B	Standard
cab44364-b35f-4cad-a0f8-b2b885873ebb	-	March 29, 2023, 20:51:57 (UTC+05:30)	175.0 B	Standard

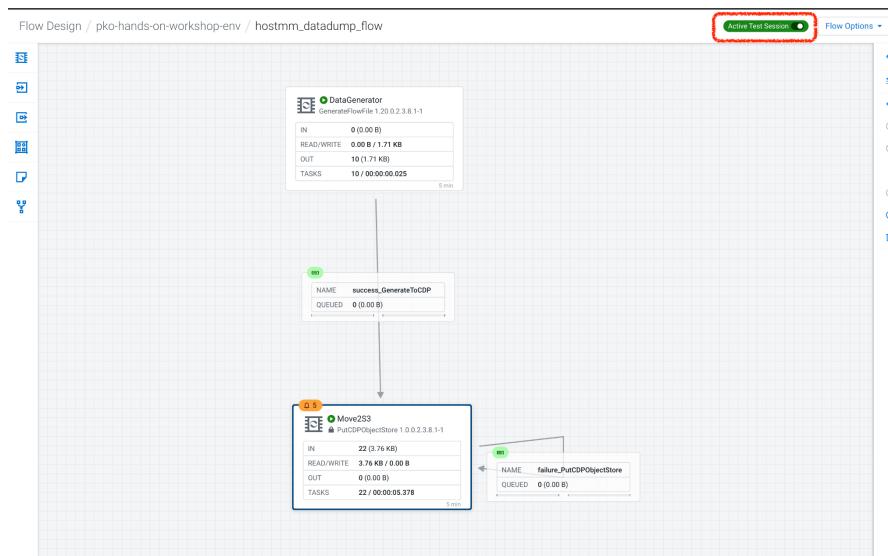
[You will not be able to access this S3 bucket by yourself but the instructor will show you where everyone's data is moving to]

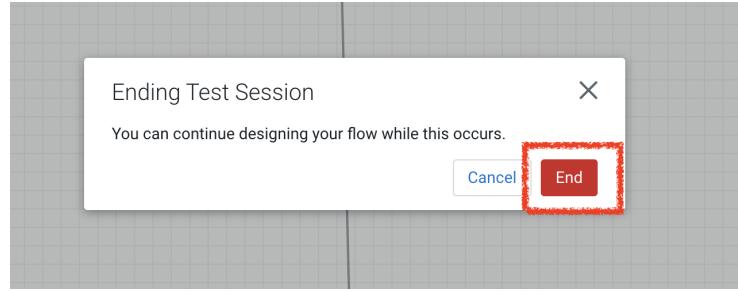
4. Move the Flow to the Flow Catalog

After the flow has been created and tested we can now PUBLISH the flow to the Flow Catalog

Step 1: STOP the current test session

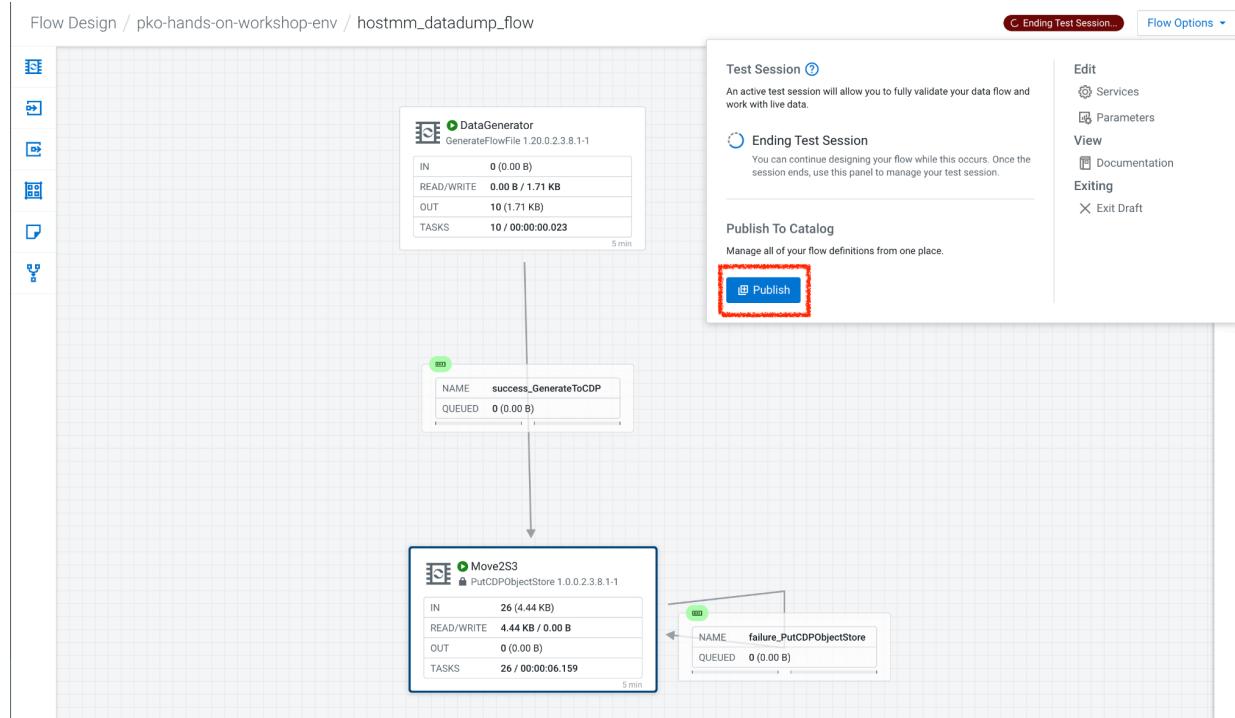
STOP the current test session by clicking on the green tab on top right and click END





Step 2: PUBLISH the flow

Once the session stops, click on **FLOW OPTION** on the top right corner of your screen and click on **PUBLISH**



Step 3: Give your flow a name and click on **PUBLISH**

Flow Name : {user_id}_datadump_flow

Custom Flow Definition

Publish A New Flow

Flow Name 12/200

Flow Description 0/1K

Version Comments 15/1K

The flow will now be visible on the **FLOW CATALOG** and is ready to be deployed

»

⌘ DataDumpFlow
Updated 3 seconds ago by Manick Mehra

Only show deployed versions

Version	Deployments	Associated Drafts
1	0	1

ASSOCIATED DRAFTS (1)
[aws pko-hands-on-workshop-env](#)
• [DataDumpFlow](#)

CRN #
[crn:cdp:df:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:flow:DataDumpFlow](#)

CREATED
2023-03-29 20:58 IST by Manick Mehra
"Initial Version"

5. Deploying the Flow

Step 1: Search for the flow in the Flow Catalog

The screenshot shows a search interface for a 'Flow Catalog'. A search bar at the top contains the text 'DataDumpFlow'. Below the search bar, a list of flows is displayed, with the first item being 'DataDumpFlow'. The interface includes a header with the title 'Flow Catalog' and a sorting option 'Name ↑'.

Click on the Flow, you should see the following:

The screenshot shows the details page for the 'DataDumpFlow' flow. At the top, there is a title 'DataDumpFlow' with a subtitle 'Updated 4 minutes ago by Manick Mehra' and an 'Actions' dropdown menu. Below this, there is a 'FLOW DESCRIPTION' section with the note 'No description specified' and a 'CRN #' field containing 'crn:cdp:df:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:flow:DataDumpFlow'. There is also a checkbox labeled 'Only show deployed versions'. A table below lists the flow's version history, showing one version (1) with 0 deployments and 1 associated draft. At the bottom, there are buttons for 'Deploy →', 'Download', and 'Create New Draft'.

Version	Deployments	Associated Drafts
1	0	1

Step 2: Deploy the flow

Click on **Version 1**, you should see a **Deploy** Option appear shortly. Then click on **Deploy**.

The screenshot shows a list of flow versions. Version 1 is selected, and a red box highlights the 'Deploy →' button. Other columns include 'Deployments' (0) and a download link.

Version	Deployments
1	0

LAST UPDATE
2021-09-23 11:52 CDT by Nasheb Ismaily
"Initial Version"

CRN #
crn:cdp:df:us-west-1:558bc1d2-8867-4357-8524-311d51259233:flow:syslog-to-kafka...

Step 3: Select the CDP environment

Select the CDP environment where this flow will be deployed and click on **CONTINUE**

NOTE: THE NAME OF THE ENVIRONMENT WILL BE SHARED BY THE INSTRUCTOR

The dialog is titled 'New Deployment'. It shows a selected flow definition ('DataDumpFlow, Version 1') and a dropdown menu for 'Select an environment'. A search bar is provided to filter environments. Three environments are listed: 'meta-workshop', 'pko-hands-on-workshop-env', and 'pse-workshop'. The 'meta-workshop' environment is currently selected.

New Deployment

Select the target environment

Selected Flow Definition

NAME	VERSION
DataDumpFlow	1

Target Environment

Select an environment

Filter by name

Name	K8s Node Allocation
aws meta-workshop	14% (3 of 21)
aws pko-hands-on-workshop-env	15% (3 of 20)
aws pse-workshop	Workspace unavailable

Step 4: Deployment Name

Give the deployment a unique name(
{user_id}_flow_prod),
then click Next.
Example :
apac01_flow_prod

Overview

Deployment Name
 (✓) Deployment name is valid

Selected Flow Definition

NAME	VERSION
DataDumpFlow	1

Target Environment

NAME
pko-hands-on-workshop-env

Click NEXT

Step 5: Set the NiFi Configuration

We can let everything be the default here and click NEXT

NiFi Configuration

NiFi Runtime Version Change Version

CURRENT VERSION
Latest Version (1.20.0.2.3.8.0-31)

ⓘ 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 →

Step 6: Set the Parameters

Set the Username, Password and the Directory name and click NEXT

CDP Workload User: apacXY

CDP Workload User Password:

ApacLabs@23

S3 Directory: dirFlowCatalogDataDump

CDP Environment : DummyParameter

[The CDP Environment parameter that shows here is used at the time we perform a test run on our test session. It holds the CDP Environment configuration resources files such as ssl-client.xml, hive-site.xml and core-site.xml. You do not have to specify these to deploy your flow from the flow catalog as it automatically picks up those files,hence we give a dummy value to this. To avoid giving a dummy value, this parameter can be deleted before we publish the flow]

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.

SHOW: Sensitive No value

hostmm_datadump_flow (4)

CDP Workload User

11/100K

host_mmehra

CDP Workload User Password

0/100K

Enter parameter values.

CDPEnvironment

10/100K

DummyValue

S3 Directory

7/100K

LabData

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

Number of NiFi Nodes

Auto Scaling

Disabled

1

Nodes

Cancel

← Previous

Next →

Step 7: Set the cluster size

Select the Extra Small size and click NEXT. In this step you can configure how your flow will autoscale, but keep it disabled for this lab.

Step 8: Add Key Performance indicators

Set up KPIs to track specific performance metrics of a deployed flow.

Click on “Add New KPI”

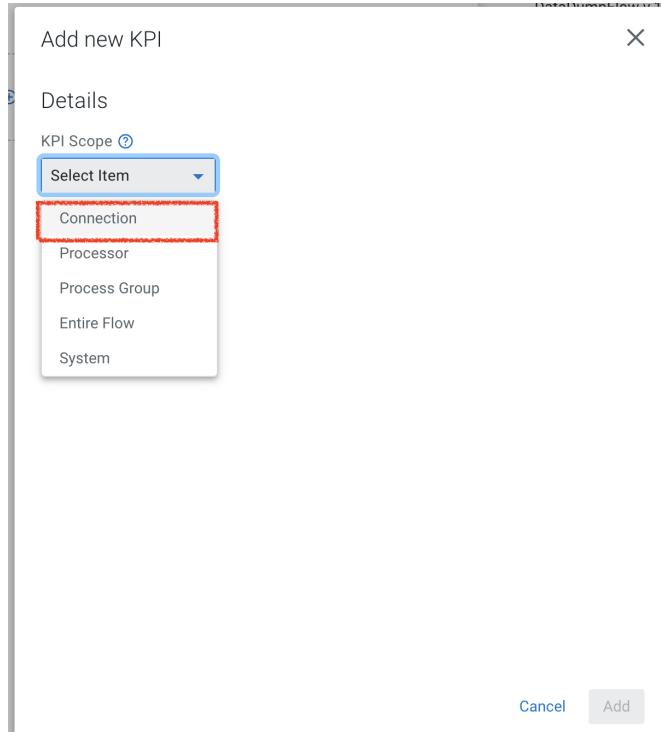
Key Performance Indicators

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](#)

In the KPI Scope drop-down list, choose “Connection”



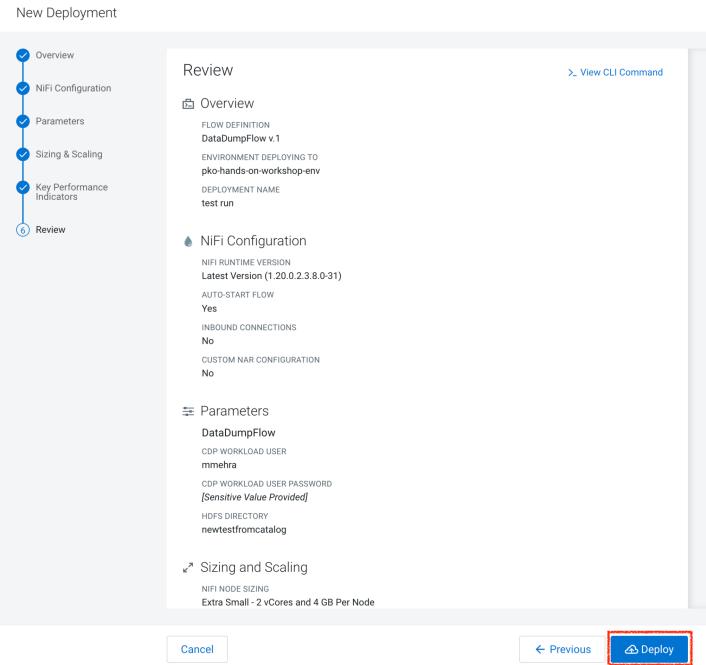
In the “Add New KPI” window, add an alert as below

The screenshot shows the 'Add new KPI' dialog box. In the 'Details' section, 'Connection Scope' is set to 'Connection' and 'Connection Name' is 'failure_Move2S3'. The 'Metric to Track' is 'Percent Full'. The 'METRIC DESCRIPTION' is 'The percentage of connection that is full'. In the 'Alerts' section, there are two options: 'Trigger alert when metric is greater than' (checked) with a value of '50' and 'Percent' selected; and 'Trigger alert when metric is less than' (unchecked) with a 'Value' input and 'Percent' selected. Below these, it says 'Alert will be triggered when metric is outside the boundary(s) for' followed by a '2 Minutes' input. At the bottom right are 'Cancel' and 'Add' buttons.

Click Add and then Click Next

The screenshot shows the 'Key Performance Indicators' page. It displays a card for a KPI named 'failure_Move2S3' with the metric 'Percent Full' set to trigger an alert if it is greater than 50% for at least 2 minutes. Below this, there is a dashed box placeholder for adding a new KPI, with the text '(+) Add New KPI'. At the bottom are 'Cancel', '< Previous', and 'Next >' buttons, where 'Next >' is highlighted with a red box.

Step 9: Click Deploy



The “Deployment Initiated” message will be displayed. Wait until the flow deployment is completed, which might take a few minutes.

»

test run
aws pko-hands-on-workshop-env

KPIs System Metrics **Alerts**

Active Alerts ?

No alerts to display.

Event History ?

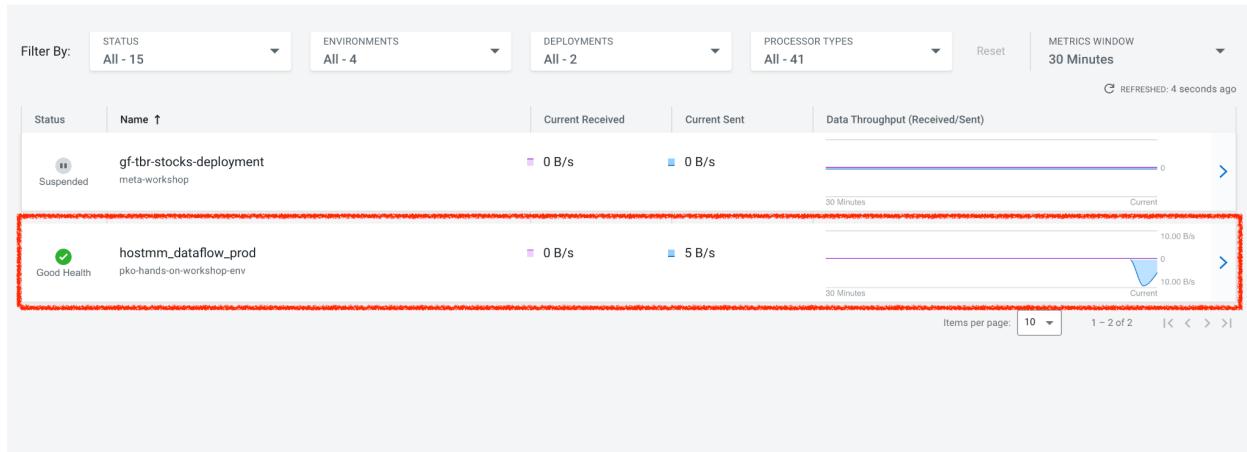
SHOW ONLY: Info Warning Error

Deployment Initiated 2023-03-29 21:20 IST v

Load More

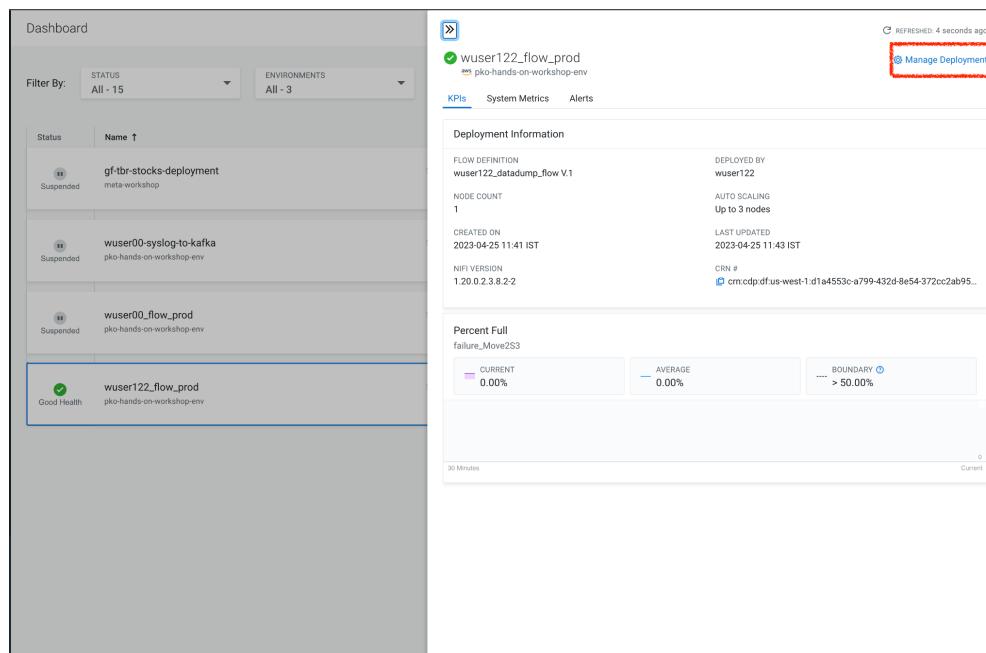
When deployed, the flow will show up on the Data flow dashboard, as below:

Dashboard



6. Viewing details of the deployed flow

Click on the flow in the Dashboard and select Manage Deployment



Step 1 : Manage KPI and Alerts

Click on the KPI tab to get the list of KPIs that have been set. You also have an option to modify or add more KPIs to your flow here.

[← Back to Deployment Details](#)

Deployment Manager

STATUS Good Health	DEPLOYMENT NAME wuser122_flow_prod	FLOW DEFINITION wuser122_datadump_flow V.1	DEPLOYED BY wuser122
NODE COUNT 1	AUTO SCALING Up to 3 nodes	CREATED ON 2023-04-25 11:41 IST	LAST UPDATED 2023-04-25 11:43 IST
aws ENVIRONMENT pko-hands-on-workshop-env	REGION Asia Pacific (Mumbai)	NIFI RUNTIME VERSION 1.20.0.2.3.8.2-2	CRN # crn:cdp:df:us-west-1:d1a4553c-a799-432d-8e54-372...

[>_ Recreate Deployment CLI Command](#)

Deployment Settings

KPIs and Alerts (highlighted with a red box) Sizing and Scaling Parameters NiFi Configuration

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](#)

Connection: failure_Move2S3
Metric to Track: Percent Full
Alert Set: Notify if greater than 50 Percent, for at least 2 minutes.

[Add New KPI](#)

Step 2 : Manage Sizing and Scaling

Click on the Sizing and Scaling tab to get detailed information

Dashboard / wuser122_flow_prod / Deployment Manager

STATUS Good Health	DEPLOYMENT NAME wuser122_flow_prod	FLOW DEFINITION wuser122_datadump_flow V.1	DEPLOYED BY wuser122
NODE COUNT 1	AUTO SCALING Up to 3 nodes	CREATED ON 2023-04-25 11:41 IST	LAST UPDATED 2023-04-25 11:43 IST
aws ENVIRONMENT pko-hands-on-workshop-env	REGION Asia Pacific (Mumbai)	NIFI RUNTIME VERSION 1.20.0.2.3.8.2-2	CRN # crn:cdp:df:us-west-1:d1a4553c-a799-432d-8e54-372...

[>_ Recreate Deployment CLI Command](#)

Deployment Settings

KPIs and Alerts (highlighted with a red box) Sizing and Scaling Parameters NiFi Configuration

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

Number of NiFi Nodes

Auto Scaling (highlighted with a red box)
Enabled
Min. Nodes: 1 Max. Nodes: 3

Step 3 : Manage Parameters

The parameters that we earlier created can be managed from the Parameters tab. Click on Parameters.

Deployment Settings

KPIs and Alerts Sizing and Scaling Parameters NiFi Configuration

Parameters

ⓘ Running Processors that are affected by the Parameter changes will automatically be restarted.

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.

SHOW: Sensitive No value

wuser122_datadump_flow (4)

CDP Workload User
wuser122

8/100K

CDP Workload User Password
Sensitive value provided.

0/100K

CDPEnvironment ⓘ
DummyValue

10/100K

Step 4 : NiFi Configurations

If you have set any configuration wrt to Nifi they will show up on the 'NiFi Configuration' tab

Deployment Settings

KPIs and Alerts Sizing and Scaling Parameters NiFi Configuration

NiFi Configuration

Inbound Connection Details

? Inbound Connection has not been configured for this deployment.

Custom NAR Configuration

? Custom NAR has not been configured for this deployment.

Step 5 : View the deployed flow in NiFi

- Select ACTIONS on the Deployment Manager page and then click on ‘View in NiFi’

Dashboard / wuser122_flow_prod / Deployment Manager

[← Back to Deployment Details](#)

Deployment Manager

aws ENVIRONMENT pk0-hands-on-workshop-env	DEPLOYMENT NAME wuser122_flow_prod	FLOW DEFINITION wuser122_datadump_flow V.1	DEPLOYED BY wuser122
STATUS Good Health	AUTO SCALING Up to 3 nodes	CREATED ON 2023-04-25 11:41 IST	LAST UPDATED 2023-04-25 11:43 IST
NODE COUNT 1	REGION Asia Pacific (Mumbai)	NIFI RUNTIME VERSION 1.20.0.2.3.8.2-2	CRN # crn:cdp:df:us-wes

[Actions](#)

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

[Recreate Deployment CLI Command](#)

Deployment Settings

KPIs and Alerts Sizing and Scaling Parameters NiFi Configuration

NiFi Configuration

Inbound Connection Details

? Inbound Connection has not been configured for this deployment.

Custom NAR Configuration

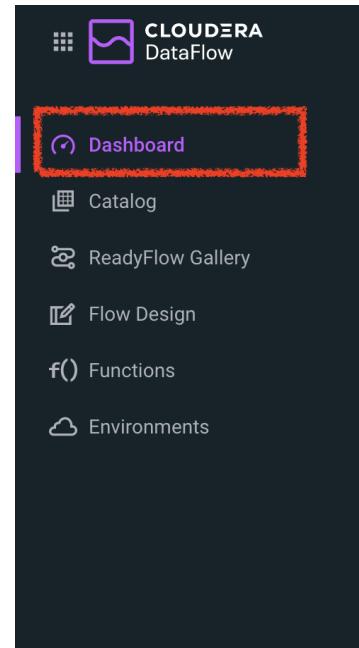
? Custom NAR has not been configured for this deployment.

This will open the flow in the NiFi UI.

The screenshot shows the Cloudera Flow Management interface. At the top, there's a toolbar with various icons for operations like start, stop, pause, and monitor. Below the toolbar, a navigation bar includes links for 'Navigate', 'Search' (with a magnifying glass icon), and 'Operate' (with a gear icon). The main workspace displays a process flow with several nodes connected by arrows. A specific node is highlighted with a blue border. To the right of the workspace, a detailed view of a flow is shown for the process group 'wuser122_flow_prod'. This view includes metrics for Queued (0 bytes), In (0 bytes → 0), Read/Write (1.71 KB / 1.71 KB), and Out (0 → 0 bytes), all updated 5 minutes ago. The bottom of the screen features a footer with status icons and a '06:28' timestamp.

Step 6 : Terminate the flow

As we have completed the Lab, it is best to terminate this flow. Follow the below given procedure to terminate your flow.



Select Dashboard from the Cloudera Data Flow UI

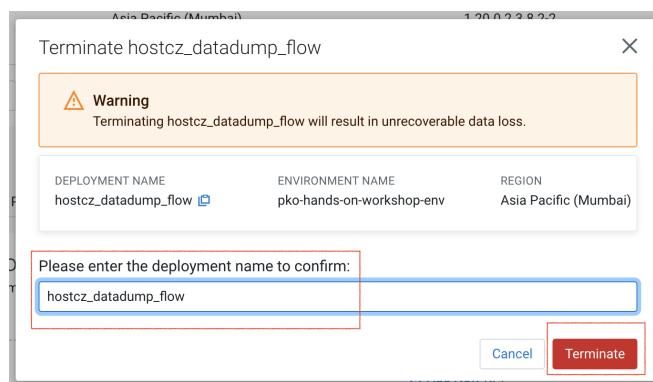
Select your flow and go to Manage Deployment

A screenshot of the Cloudera DataFlow UI. On the left is a sidebar with "Dashboard" (highlighted with a red box), "Catalog", "ReadyFlow Gallery", "Flow Design", "Functions", and "Environments". The main area is titled "Dashboard" and shows a list of flows. One flow, "hostcz_datadump_flow" (status: Deploying), is highlighted with a red box. To the right of the flow list is a detailed view for "hostcz_datadump_flow" in "pko-hands-on-workshop-env". It includes sections for "KPIs", "System Metrics", and "Alerts" (which is underlined). Below these is an "Event History" table with the following data:

On the Deployment Manager Page, Select **Actions** and click on **Terminate**

The screenshot shows the Cloudera DataFlow Deployment Manager interface. On the left, there's a sidebar with options like Dashboard, Catalog, ReadyFlow Gallery, Flow Design, Functions, and Environments. The main area displays deployment details for 'hostcz_datadump_flow'. It includes sections for STATUS (Good Health), DEPLOYMENT NAME, FLOW DEFINITION, NODE COUNT (1), ENVIRONMENT (pko-hands-on-workshop-env), AUTO SCALING (Disabled), REGION (Asia Pacific (Mumbai)), NiFi RUNTIME VERSION (1.20.0.2.3.8.2.2), and DEPLOYED BY (Manick Mehra). A timestamp indicates the page was refreshed 12 seconds ago. On the right, there's an 'Actions' dropdown menu with options: View in NiFi, Suspend Flow, Change NiFi Runtime Version, Restart Deployment, and Terminate. The 'Terminate' option is highlighted with a red box.

In the next dialog box, enter the name of the flow we are trying to terminate and click on **Terminate**



You will now see that the termination process has started.

The screenshot shows the Cloudera DataFlow Alerts page. It features tabs for KPIs, System Metrics, and Alerts, with 'Alerts' selected. Under 'Active Alerts', it says 'No alerts to display.' In the 'Event History' section, there's a table of log entries. One entry, 'Deployment Termination Initiated' at 2023-04-30 00:04 IST, is highlighted with a red box. Other entries include 'Deployment Successful' at 2023-04-29 23:58 IST, 'NiFi Flow Started' at 2023-04-29 23:58 IST, and 'KPI Alert Rules Activated' at 2023-04-29 23:58 IST. A 'Load More' button is at the bottom.

Lab 2 : Migrating Existing Data Flows to CDF-PC

1. Overview

The purpose of this workshop is to demonstrate how existing NiFi flows can be migrated to the Data Flow Experience. This workshop will leverage an existing NiFi flow template that has been designed with the best practices for CDF-PC flow deployment.

The existing NiFi Flow will perform the following actions:

1. Generate random syslogs in 5424 Format
2. convert the incoming data to a JSON using record writers
3. Apply a SQL filter to the JSON records
4. Send the transformed syslog messages to Kafka

Note that a parameter context has already been defined in the flow and the queues have been uniquely named.

For this we will be leveraging the DataHubs which have already been created, namely:

- workshop-asean-streams-mess
- workshop-asean-stream-analytics

2. Pre-requisites

2.1. Create a Kafka Topic

1. Login to Streams Messaging Manager by clicking the appropriate hyperlink in the Streams Messaging Datahub (workshop-asean-streams-mess)

The screenshot shows the Cloudera Management Console interface. On the left, the navigation sidebar includes options like Dashboard, Environments, Data Lakes, User Management, Data Hub Clusters, Data Warehouses, ML Workspaces, Classic Clusters, Audit, Shared Resources, and Global Settings. The main content area displays the 'Environments / pko-hands-on-workshop-env / Clusters' page. A cluster named 'pko-hands-on-workshop-env' is selected, showing its details: Name: pko-workshop-dl, Nodes: 2, Status: Running, Status Reason: DataLake is running, Scale: Light Duty, CRN: crn:cdp:datalake:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:datalake:6a9aed14-c21b-42fe-88c4-e9f7e8a517a0. Below this, a table lists other Data Hubs: sbb-analytics-cluster (CDH 7.2.16), nifi-flow-mgmt-cluster (CDH 7.2.16), and kafka-smm-cluster (CDH 7.2.16).

Below the clusters section, there's a 'Services' section with links to CM UI, Schema Registry, and Streams Messaging Manager. The 'Streams Messaging Manager' link is highlighted with a red box.

2. Click on Topics in the left tab

The screenshot shows the 'Overview' page of the Streams Messaging Manager. At the top, it displays 'Producers: 11' and 'Brokers: 3'. Below this, a table lists topics: '_consumer_offsets', '_CruiseControlMetrics', '_KafkaCruiseControlModelTrainingSamples', '_KafkaCruiseControlPartitionMetricSamples', '_smm_alert_notifications', and '_smm_consumer_metrics'. The 'Topics' tab in the left sidebar is highlighted with a red box.

4. Create a Topic with the following parameters then click **Save**:

- **Name:** <username>-syslog
- **Partitions:** 1
- **Availability:** Moderate
- **Cleanup Policy:** Delete

Add Topic

TOPIC NAME	PARTITIONS
workshop001-syslog	1

Availability

MAXIMUM	HIGH	MODERATE	LOW	CUSTOM

REPLICATION FACTOR 3 MIN INSYNC REPLICA 2	REPLICATION FACTOR 3 MIN INSYNC REPLICA 1	REPLICATION FACTOR 2 MIN INSYNC REPLICA 1	REPLICATION FACTOR 1 MIN INSYNC REPLICA 1
--	--	--	--

Limits

CLEANUP.POLICY

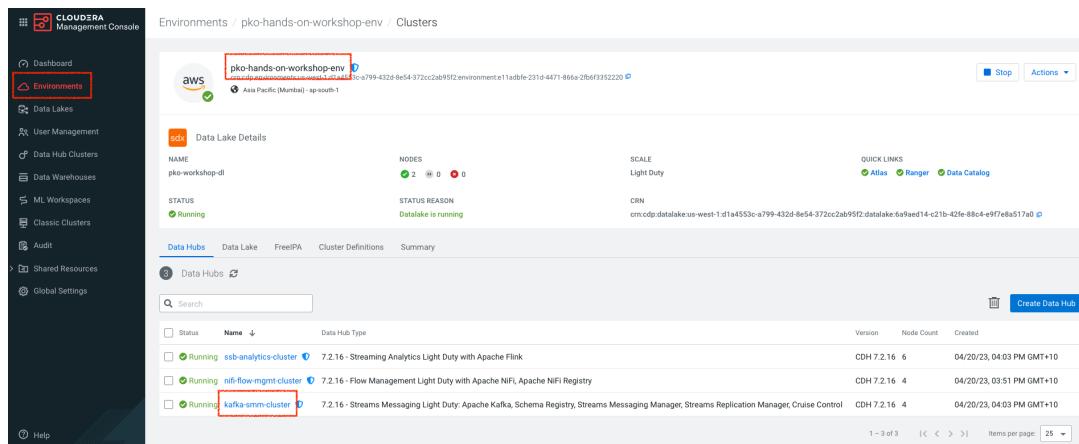
delete

[Advanced](#) [Cancel](#) **Save**

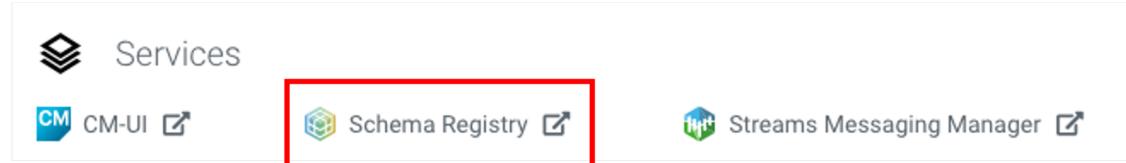
Note: The Flow will not work if you set the Cleanup Policy to anything other than **Delete**. This is because we are not specifying keys when writing to Kafka.

2.2. Create a Schema in Schema Registry

1. Login to Schema Registry by clicking the appropriate hyperlink in the Streams Messaging Datahub(kafka-smm-cluster)



Status	Name	Type	Version	Node Count	Created
Running	ssb-analytics-cluster	7.2.16 - Streaming Analytics Light Duty with Apache Flink	CDH 7.2.16	6	04/20/23, 04:03 PM GMT+10
Running	nifi-flow-mgmt-cluster	7.2.16 - Flow Management Light Duty with Apache NiFi, Apache NiFi Registry	CDH 7.2.16	4	04/20/23, 03:51 PM GMT+10
Running	kafka-smm-cluster	7.2.16 - Streams Messaging Light Duty: Apache Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control	CDH 7.2.16	4	04/20/23, 04:03 PM GMT+10



2. Click on the + button on the top right to create a new schema.



- **Name:** <username>-syslog
- **Description:** syslog schema for dataflow workshop
- **Type:** Avro schema provider
- **Schema Group:** Kafka
- **Compatibility:** Backward
- **Evolve:** True
- **Schema Text:** Copy and paste the schema text below into the “Schema Text” field

```
{
  "name": "syslog",
  "type": "record",
  "namespace": "com.cloudera",
  "fields": [
    {
      "name": "priority",
      "type": "int"
    },
    {
      "name": "severity",
      "type": "int"
    },
    {
      "name": "facility",
      "type": "int"
    },
    {
      "name": "version",
      "type": "int"
    },
    {
      "name": "timestamp",
      "type": "long"
    },
    {
      "name": "hostname",
      "type": "string"
    },
    {
      "name": "body",
      "type": "string"
    },
    {
      "name": "appName",
      "type": "string"
    }
  ]
}
```

```
{  
    "name": "procid",  
    "type": "string"  
},  
{  
    "name": "messageid",  
    "type": "string"  
},  
{  
    "name": "structuredData",  
    "type": {  
        "name": "structuredData",  
        "type": "record",  
        "fields": [  
            {  
                "name": "SDID",  
                "type": {  
                    "name": "SDID",  
                    "type": "record",  
                    "fields": [  
                        {  
                            "name": "eventId",  
                            "type": "string"  
                        },  
                        {  
                            "name": "eventSource",  
                            "type": "string"  
                        },  
                        {  
                            "name": "iut",  
                            "type": "string"  
                        }  
                    ]  
                }  
            }  
        ]  
    }  
},  
]
```

Note: The name of the Kafka Topic you previously created and the Schema Name must be the same.

Click on **SAVE**.

Add New Schema

NAME *

DESCRIPTION *

TYPE *

SCHEMA GROUP *

COMPATIBILITY

 EVOLVE

SCHEMA TEXT *

```
61 },
62 {
63   "name": "eventSource",
64   "type": "string"
65 },
66 {
67   "name": "iut",
68   "type": "string"
69 }
70 ]
71 }
72 ]
73 }
74 }
75 ]
76 }
77 }
```


SCHEMA REGISTRY All Schemas

Search by name Q Sort: Last Updated ▾

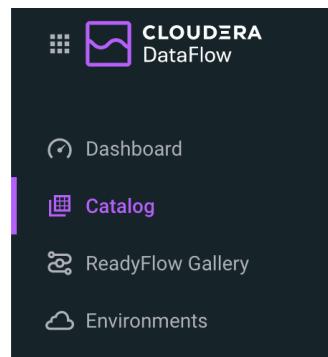
workshop001-syslog	TYPE avro	GROUP Kafka	BRANCH 1 ↗	SERIALIZER & DESERIALIZER 0
 View				

[New Schema](#)

Lab 3 : Operationalizing Externally Developed Data Flows with CDF-PC

1. Import the Flow into the CDF-PC Catalog

- Open the CDF-PC data service and click on Catalog in the left tab.



- Select Import Flow Definition on the Top Right

+ Import Flow Definition

- Add the following information:

- Flow Name:** <username>-syslog-to-kafka
- Flow Description:**

Reads Syslog in RFC 5424 format, applies a SQL filter, transforms the data into JSON records, and publishes to Kafka

- NiFi Flow Configuration:** syslog-to-kafka.json (From the resources downloaded earlier)
- Version Comments:** Initial Version

Import Flow Definition X

Flow Name	syslog-to-kafka
Flow Description	120/1000 Generates Syslog in RFC 5424 format, applies a SQL filter, transforms the data into JSON records, and publishes to Kafka
NiFi Flow Configuration	syslog-to-kafka.json (✓)
Version Comments	15/1000 Initial Version

[Cancel](#) [Import](#)

Click **IMPORT**

2. Deploy the Flow in CDF-PC

1. Search for the flow in the Flow Catalog

Flow Catalog

Import Flow Definition

REFRESHED: 20 seconds ago

Name ↑	Type	Versions	Last Updated	>
syslog-to-kafka	Custom Flow Definition	1	3 minutes ago	>

2. Click on the Flow, you should see the following:

» REFRESHED: 7 seconds ago

syslog-to-kafka Updated 4 minutes ago by Nasheb Ismaily Actions ▾

FLOW DESCRIPTION

Generates Syslog in RFC 5424 format, applies a SQL filter, transforms the data into JSON records, and publishes to Kafka

CRN # [crn:cdp:df:us-west-1:558bc1d2-8867-4357-8524-311d51259233:flow:syslog-to-kafka](#)

Only show deployed versions

Version	Deployments
1	0

3. Click on **Version 1**, you should see a **Deploy** Option appear shortly. Then click on **Deploy**.

The screenshot shows a deployment history table with one entry. The first column is 'Version' with value '1'. The second column is 'Deployments' with value '0'. Below the table is a blue button labeled 'Deploy →'. To the right of the button is a 'Download' link.

Version	Deployments
1	0

LAST UPDATE
2021-09-23 11:52 CDT by Nasheb Ismaily
"Initial Version"

CRN #
[crn:cdp:df:us-west-1:558bc1d2-8867-4357-8524-311d51259233:flow:syslog-to-kafka...](#)

4. Select the CDP environment where this flow will be deployed, then click **Continue**.

NOTE: THE NAME OF THE ENVIRONMENT WILL BE SHARED BY THE INSTRUCTOR

New Deployment X

Select the target environment

ⓘ Sensitive data never leaves the environment. Changing the environment after this step requires restarting the deployment process.

Selected Flow Definition

	NAME workshop001-syslog-to-kafka	VERSION 1
--	-------------------------------------	--------------

Target Environment

workshop-asean 15% (3 of 20) ▾

Cancel Continue →

5. Give the deployment a unique name, then click **Next**.

Example : {user_id}-syslog-to-kafka

New Deployment

The screenshot shows the 'New Deployment' process with six steps. Step 1, 'Overview', is currently selected and highlighted in blue. The other steps are numbered 2 through 6: NiFi Configuration, Parameters, Sizing & Scaling, Key Performance Indicators, and Review.

Overview

Deployment Name: workshop001-syslog-to-kafka

Deployment name is valid

Selected Flow Definition:

NAME	VERSION
workshop001-syslog-to-kafka	1

Target Environment:

NAME
aws workshop-asean

6. In the NiFi Configuration screen, click **Next**.

The screenshot shows the 'NiFi Configuration' screen, which is the second step in the deployment process. The left sidebar shows the current step is 'NiFi Configuration' (step 2), with other steps 3 through 6 listed below it.

NiFi Configuration

NiFi Runtime Version:

CURRENT VERSION
Latest Version (1.20.0.2.3.8.1-1)

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

Buttons at the bottom:

- Cancel
- ← Previous
- Next → (This button is highlighted with a red box)

7. Add the Flow Parameters as below, then click **Next**.

- **CDP Workload User** - The workload username for the current user
 - Example : workshop001
- **CDP Workload Password** - The workload password for the current user
P@ssw0rd@2023
- **Filtre Rule** - SELECT * FROM FLOWFILE
- **Kafka Broker Endpoint** - A comma separated list of Kafka Brokers.
[Obtained in Lab 0, section 4]
Example: workshop-asean-streams-mess-corebroker2.workshop.dp5i-5vkq.cloudera.site:9093,workshop-asean-streams-mess-corebroker0.workshop.dp5i-5vkq.cloudera.site:9093,workshop-asean-streams-mess-corebroker1.workshop.dp5i-5vkq.cloudera.site:9093
- **Kafka Destination Topic** - <username>-syslog (Ex: workshop001-syslog)
- **Kafka Producer ID** - nifi_dfx_p1
- **Schema Name** - <username>-syslog (Ex: workshop001-syslog)
- **Schema Registry Hostname** - The hostname of the master server in the Kafka Datahub(kafka-smm-cluster)[Refer screenshot below]

The screenshot shows the Cloudera Manager interface for the 'kafka-smm-cluster' node. The node details are as follows:

- Name:** kafka-smm-cluster
- Status:** Running
- Nodes:** 4 (0 up, 0 down)
- Created At:** 04/20/23, 11:33 AM GMT+5:30
- Cluster Template:** 7.2.16 - Streams Messaging Light Duty: Apache Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control
- Reason:** Cluster started.
- AWS Environment Details:**
 - NAME:** pko-hands-on-workshop-env
 - DATA LAKE:** pko-workshop-dl
 - CREDENTIAL:** pko-hands-on-workshop-cred
 - REGION:** ap-south-1
 - AVAILABILITY ZONE:** N/A
- Services:** CM UI, Schema Registry, Streams Messaging Manager, Token Integration
- Cloudera Manager Info:**
 - CM URL:** https://kafka-asean-cluster-gateway.pko-hand.dp5i-5vkq.cloudera.site/kafka-smm-cluster/cdp-proxy/cm/home/
 - CM VERSION:** 7.9.0
 - RUNTIME VERSION:** 7.2.16-1.cdh7.2.16.p2.38683602
 - LOGS:** Command logs, Service logs
- Master:** Instance ID: 10db4b6eff7be50080, Status: Running, Public IP: 10.10.220.8, CM Server

Example: workshop-asean-streams-mess-master0.workshop.dp5i-5vkq.cloudera.site

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.

i 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

syslog-to-kafka (8)

CDP Workload User ?	11/100K
workshop001	
CDP Workload User Password ?	13/100K
.....	
Filter Rule ?	22/100K
SELECT * FROM FLOWFILE	
Kafka Broker Endpoint ?	233/100K
workshop-asean-streams-mess-corebroker2.workshop.dp5i-5vkq.cloudera.site:9093,workshop-asean-streams-mess-corebroker0.workshop.dp5i-5vkq.cloudera.site:9093,workshop-asean-streams-mess-corebroker1.workshop.dp5i-5vkq.cloudera.site:9093	
Kafka Destination Topic ?	18/100K
workshop001-syslog	
Kafka Producer ID ?	11/100K
nifi_dfx_p1	
Schema Name ?	18/100K
workshop001-syslog	
Schema Registry Hostname ?	68/100K
workshop-asean-streams-mess-master0.workshop.dp5i-5vkq.cloudera.site	

8. On the next page, define the Sizing and Scaling as follows, then click **Next**.

- **Size:** Extra Small
- **Enable Auto Scaling:** True
- **Min Nodes:** 1
- **Max Nodes:** 3

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

NiFi Node Sizing



Number of NiFi Nodes

Auto Scaling

Enabled

Min. Nodes - Max. Nodes

9. Skip the KPI page by clicking **Next** and Review your deployment. Then Click **Deploy**.

Review

View CLI Command

Overview

FLOW DEFINITION: mmehra_test v.1
ENVIRONMENT DEPLOYING TO: pko-hands-on-workshop-env
DEPLOYMENT NAME: mmehra_test

NiFi Configuration

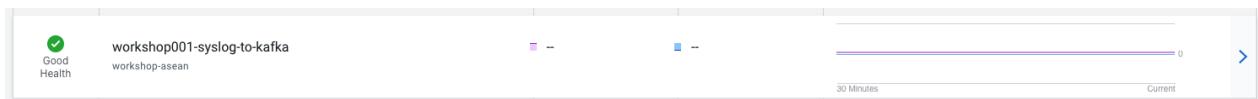
NIFI RUNTIME VERSION: Latest Version (1.20.0.2.3.8.1-1)
AUTO-START FLOW: Yes
INBOUND CONNECTIONS: No
CUSTOM NAR CONFIGURATION: No

Parameters

syslog-to-kafka
CDP WORKLOAD USER: host_mmehra
CDP WORKLOAD USER PASSWORD: [Sensitive Value Provided]
FILTER RULE: SELECT * FROM FLOWFILE
KAFKA BROKER ENDPOINT: kafka-smm-cluster-corebroker1.pko-hand.dp5i-5vkq.cloudera.site:9093,kafka-smm-cluster-corebroker0.pko-hand.dp5i-5vkq.cloudera.site:9093,kafka-smm-cluster-corebroker2.pko-hand.dp5i-5vkq.cloudera.site:9093
KAFKA DESTINATION TOPIC: mmehra_test

Cancel Previous Deploy

10. Proceed to the CDF-PC Dashboard and wait for your flow deployment to complete, which might take a few minutes. A Green Check Mark will appear once complete, which might take a few minutes.



view your flow in NiFi.

The screenshot shows the NiFi UI dashboard. On the left, a list of data flows is displayed, including:

- gf-tbr-stocks-deployment
- hostcz-syslog-to-kafka (selected)
- hostcz_flow_prod
- josua_stock_dataflow
- nifi-testimmersionday
- sai-test

On the right, detailed information for the selected flow is shown:

- KPIs**: System Metrics, Alerts
- REFRESHED: 3 seconds ago**
- hostcz-syslog-to-kafka** (Flow Definition)
hostcz-syslog-to-kafka V.1
- ENVIRONMENTS**: pko-hands-on-workshop-env
- Deployment Information**
 - DEPLOYED BY: host_cz
 - NODE COUNT: 2
 - AUTO SCALING: Up to 3 nodes
 - CREATED ON: 2023-04-27 08:21 IST
 - LAST UPDATED: 2023-04-27 20:11 IST
 - NIFI VERSION: 1.20.0.2.3.8.2-2
 - CRN #: crn:cdf:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95..
- No KPIs to display.** Set up key performance indicators to track specific aspects of your data flow to ensure it's operating as expected.
[Learn more](#)

Now click on **ACTIONS** and select **View in NiFi**

[← Back to Deployment Details](#)

Deployment Manager

STATUS Good Health	FLOW DEFINITION hostcz-syslog-to-kafka V.1	DEPLOYED BY host_cz
NODE COUNT 2	CREATED ON 2023-04-27 08:21 IST	LAST UPDATED 2023-04-27 20:11 IST
ENVIRONMENT aws pkc-hands-on-workshop-env	AUTO SCALING Up to 3 nodes	NIFI RUNTIME VERSION 1.2.0.0.2.3.8.2.2
REGION Asia Pacific (Mumbai)		CRN # crm:cdp:dfus-wes

[Actions ▾](#)

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

[Recreate Deployment CLI Command](#)

Deployment Settings

KPIs and Alerts Sizing and Scaling Parameters NiFi Configuration

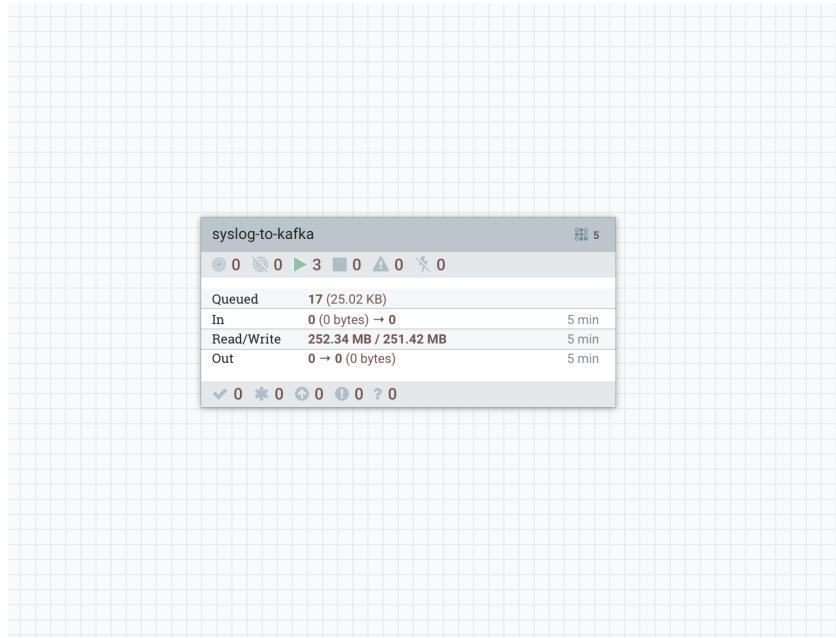
Key Performance Indicators

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

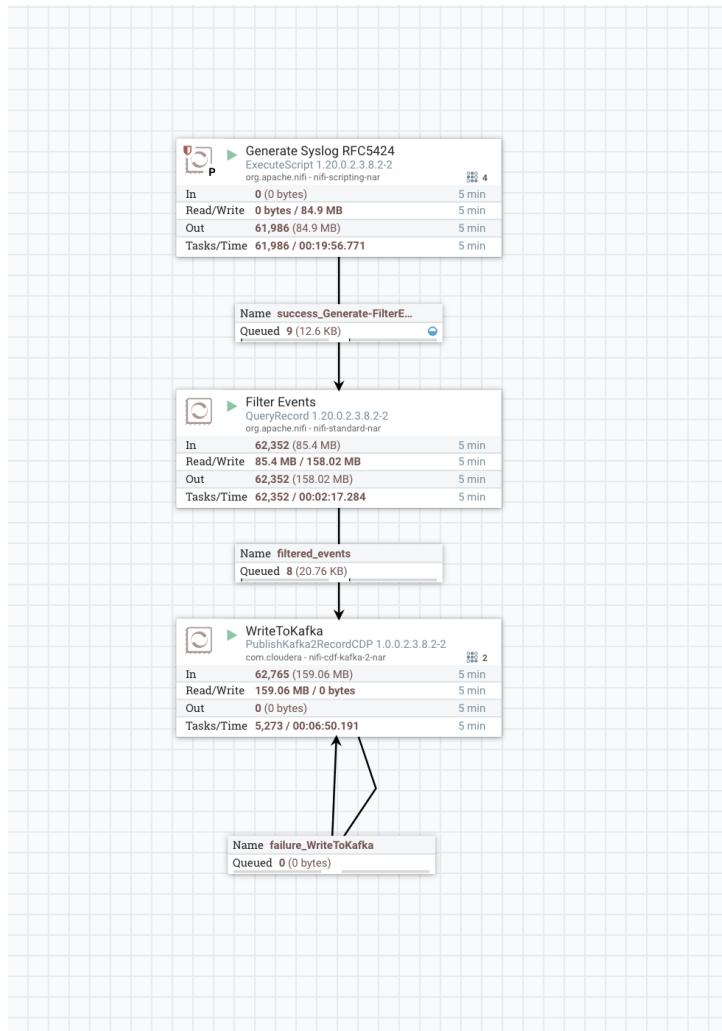
[Learn more ↗](#)

[Add New KPI](#)

The flow that you just deployed will look something like this on NiFi



Double click on the Process Group to see the flow



Lab 4 : SQL Stream Builder

1. Overview

The purpose of this workshop is to demonstrate streaming analytic capabilities using Cloudera SQL Stream Builder. We will leverage the NiFi Flow deployed in CDF-PC from the previous lab and demonstrate how to query live data and subsequently sink it to another location. The SQL query will leverage the existing syslog schema in Schema Registry.

2. Creating a Project

Step 1: Go to the SQL Stream Builder UI

SSB Interface can be reached from the Data Hub that is running the Streams Analytics, in our case - `ssb-analytics-cluster`

Within the Data Hub, click on **Streaming SQL Console**

The screenshot shows the Data Hub interface for the `ssb-analytics-cluster` cluster. At the top, there's a navigation bar with links for Data Hubs, `ssb-analytics-cluster`, and Event History. Below the navigation is a cluster summary card for `ssb-analytics-cluster`. The card includes status information (Running, 6 nodes), creation date (04/20/23, 11:33 AM GMT+5:30), cluster template (7.2.16 - Streaming Analytics Light Duty with Apache Flink), and a status reason (Cluster started). Below the summary card, there are sections for Environment Details (aws, NAME: pko-hands-on-workshop-env, DATA LAKE: pko-workshop-dl, CREDENTIAL: pko-hands-on-workshop-cred, REGION: ap-south-1, AVAILABILITY ZONE: N/A) and Services. The Services section lists CM-UI, Flink Dashboard, Job History Server, Name Node, Queue Manager, Resource Manager, Streaming SQL Console, Token Integration, and Cloudera Manager Info. The `Streaming SQL Console` link is highlighted with a red box. At the bottom, there are links for Event History, Autoscale, Endpoints (5), Tasks (4), Nodes, Network, Load Balancers, Telemetry, Repository Details, Image Details, Recipes (0), Cloud Storage, Database, and Unimgrade.

Step 2: Creation of a Project

Create a SSB Project by clicking “**New Project**” using the following details and click “**Create**”

Name : {user-id}_hol_workshop

Description : SSB Project to analyze streaming data

Create Project

Name *

Description

Override Materialized View Table Name Prefix

Source Settings

Clone URL

Branch

Allow deletions on import

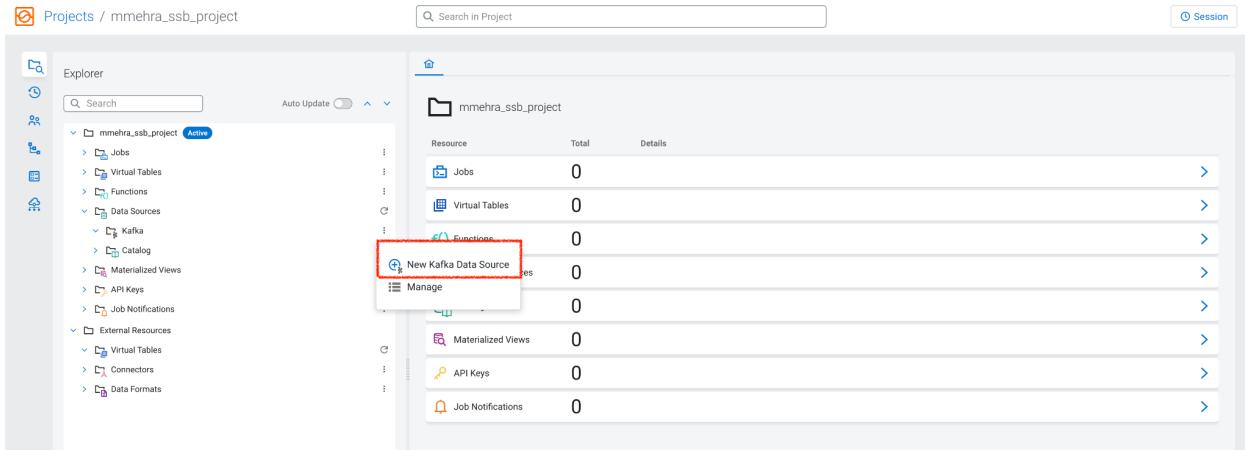
Authentication

Create

Switch to the created project. Click on **Switch**

Step 3 : Create Kafka Data Store

Create Kafka Data Store by selecting “**Data Sources**” in the left pane, clicking on the three-dotted icon next to “**Kafka**”, then selecting “**New Kafka Data Source**”.



Name : {user-id}_cdp_kafka

Brokers (Comma-separated List)

workshop-asean-streams-mess-corebroker2.workshop.dp5i-5vkq.cloudera.site:9093,workshop-asean-streams-mess-corebroker0.workshop.dp5i-5vkq.cloudera.site:9093,workshop-asean-streams-mess-corebroker1.workshop.dp5i-5vkq.cloudera.site:9093

Protocol : SASL/SSL

SASL Username : <workload-username>

Example : workshop001

SASL Password : <Set in Lab

0 Section 3> P@ssw0rd@2023

SASL Mechanism : PLAIN

The dialog has the following fields:

- Name ***: njay-demo-ssb-kafka-ds
- Brokers (Comma-separated List) ***: niay-demo-kafka-corebroker2.njay-dem.a465-9q4k.cloudera.site:9093,niay-demo-kafka-corebroker1.njay-dem.a465-9q4k.cloudera.site:9093,niay-demo-kafka-corebroker0.njay-dem.a465-9q4k.cloudera.site:9093 (highlighted with a red arrow)
- Protocol ***: SASL/SSL (highlighted with a red arrow)
- Kafka TrustStore (Optional)**: /var/lib/cloudera-scm-agent/agent-cert/cm-auto-global_truststore.jks
- Kafka TrustStore Password (Optional)**: (empty)
- Kafka KeyStore (Optional)**: (empty)
- Kafka KeyStore Password (Optional)**: (empty)
- SASL Mechanism**: PLAIN (highlighted with a red arrow)

At the bottom right are 'Cancel' and 'Create' buttons.

SASL Mechanism

SASL Username

SASL Password

Click on **VALIDATE** to test the connections once successful click on **CREATE**



Step 4: Create Kafka Table

Create Kafka Table, by selecting “**Virtual Tables**” in the left pane, clicking on the three-dotted icon next to it, then clicking on “**New Kafka Table**”.

The screenshot shows the Data Pipeline interface with the following details:

- Top Bar:** Projects / mmehra_ssbb_project, Search in Project.
- Left Sidebar (Explorer):**
 - mmehra_ssbb_project (Active)
 - Jobs
 - Virtual Tables** (highlighted with a red box)
 - Functions
 - Data Sources (1)
 - Kafka (1)
 - CDP Kafka
 - Catalog
 - Materialized Views
 - API Keys
 - Job Notifications
 - External Resources
 - Virtual Tables
 - Connectors
 - Data Formats
- Right Panel (Resource):**
 - mmehra_ssbb_proj
 - Resource
 - Jobs
 - New Kafka Table** (highlighted with a red box)
 - New Webhook Table
 - Manage
 - Catalogs
 - Materialized Views
 - API Keys
 - Job Notifications

Step 5: Configure the Kafka Table

1. Enter the following details in the Kafka Table dialog box:
 - Table Name: **{user-id}_syslog_data**
 - Kafka Cluster: <select the Kafka data source you created previously>
 - Data Format: **JSON**
 - Topic Name: <select the topic created in Schema Registry>

Kafka Table

Table Name *	workshop001_syslog_data
Kafka Cluster *	workshop001_cdp_kafka
Data Format *	JSON
Topic Name *	workshop001-syslog

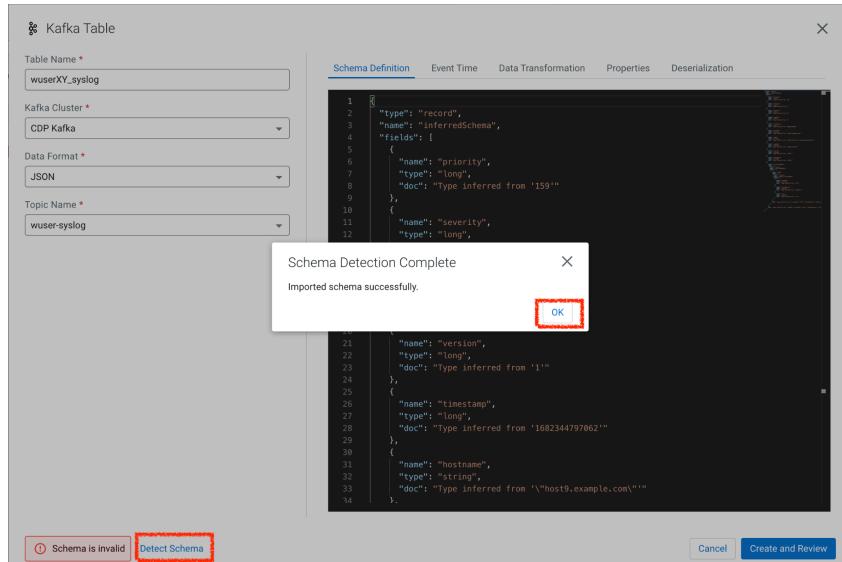
Schema is valid Detect Schema

2. When you select Data Format as AVRO, you must provide the correct Schema Definition when creating the table for SSB to be able to successfully process the topic data.

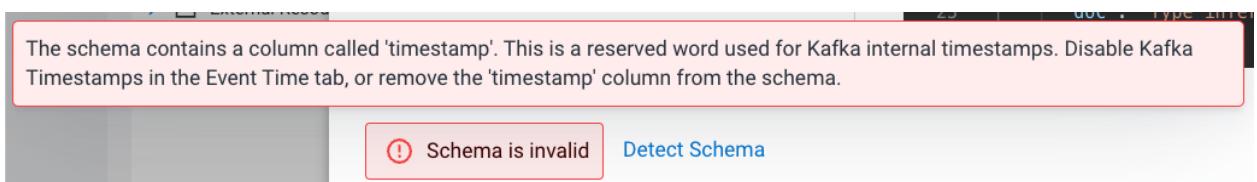
For JSON tables, though, SSB can look at the data flowing through the topic and try to infer the schema automatically, which is quite handy at times. Obviously, there must be data in the topic already for this feature to work correctly.

Note: SSB tries its best to infer the schema correctly, but this is not always possible and sometimes data types are inferred incorrectly. You should always review the inferred schemas to check if it's correctly inferred and make the necessary adjustments.

Since you are reading data from a JSON topic, go ahead and click on **Detect Schema** to get the schema inferred. You should see the schema be updated in the **Schema Definition** tab.



3. You will also notice that a "Schema is invalid" message appears upon the schema detection. If you hover the mouse over the message it shows the reason:



You will fix this in the next step.

4. Each record read from Kafka by SSB has an associated timestamp column of data type **TIMESTAMP ROWTIME**. By default, this timestamp is sourced from the internal timestamp of the Kafka message and is exposed through a column called **eventTimestamp**.

However, if your message payload already contains a timestamp associated with the event (**event time**), you may want to use that instead of the Kafka internal timestamp.

In this case, the syslog message has a field called **"timestamp"** that contains the timestamp you should use. You want to expose this field as the table's **"event_time"** column. To do this, click on the Event Time tab and enter the following properties:

- Use Kafka Timestamps: **Disable**
- Input Timestamp Column: **timestamp**
- Event Time Column: **event_time**
- Watermark Seconds: **3**

Kafka Table

Table Name *	syslog_data
Kafka Cluster *	dh-kafka
Data Format *	JSON
Topic Name *	araujo-syslog-json
<input style="width: 100px; margin-bottom: 5px;" type="button" value="Schema Definition"/> <input style="width: 100px; margin-bottom: 5px;" type="button" value="Event Time"/> <input style="width: 100px; margin-bottom: 5px;" type="button" value="Data Transformation"/> <input style="width: 100px; margin-bottom: 5px;" type="button" value="Properties"/>	
Input Timestamp Column	
timestamp	
Event Time Column	
event_time	
Watermark Seconds	
3	
<input type="checkbox"/> Use Kafka Timestamps	

5. Now that you have configured the event time column, click on **Detect Schema** again. You should see the schema turn valid:

 Schema is valid

6. Click the **Create and Review** button to create the table.

Kafka Table

```

CREATE TABLE `ssb`.`host_test`.`wuser_syslog_kafka` (
  `id` INT,
  `level` INT,
  `severity` INT,
  `facility` INT,
  `version` INT,
  `time` BIGINT,
  `hostname` VARCHAR(2147483647),
  `body` VARCHAR(2147483647),
  `appname` VARCHAR(2147483647),
  `process` VARCHAR(2147483647),
  `messageid` VARCHAR(2147483647),
  `structuredData` ROW<'$ID' ROW<'eventId' VARCHAR(2147483647), 'eventSource' VARCHAR(2147483647), 'sut' VARCHAR(2147483647)>,
  `event_time` AS TO_TIMESTAMP_LTZ(`timestamp`, 3),
  `watermark` AS `event_time` - INTERVAL '3' SECOND
) WITH (
  `properties.security.protocol` = 'SASL_SSL',
  `scm.startup.mode` = 'earliest-offset',
  `properties.sasl.jaas.config` = "org.apache.kafka.common.security.plain.PlainLoginModule required username=\"wuser\$1\" password=\"*****\"",
  `properties.sasl.kerberos.principal` = "wuser@CLUSTERNAME.COM",
  `properties.ssl.truststore.location` = '/var/lib/cloudera-scm-agent/gagent-cert/cn-auto-global_truststore.jks',
  `properties.auto.offset.reset` = 'earliest',
  `properties.sasl.mechanism` = 'PLAIN',
  `format` = 'json',
  `properties.bootstrap.servers` = 'kafka-sm-cluster--corebroker1,pko-hand-dp51-5vkq.cloudera.site:9093,kafka-sm-cluster--corebroker0,pko-hs',
  `connector` = 'kafka',
  `properties.transaction.timeout.ms` = '900000',
  `topic` = 'syslog_test'
)

```

 Success
wuser_syslog_kafka has been saved

Review the table's DDL and click **Close**.

Step 6: Create a Flink Job

Create a Flink Job, by selecting “**Jobs**” in the left pane, clicking on the three-dotted icon next to it, then clicking on “**New Job**”.

Projects / mmehra_ss_project

Search in Project

Explorer

Auto Update

mmehra_ss_project (Active)

Jobs (1)

Virtual Tables

Functions

Data Sources

Materialized Views

API Keys

Job Notifications

External Resources

New Job

Jobs (0)

Virtual Tables (1) kafka

Functions (0)

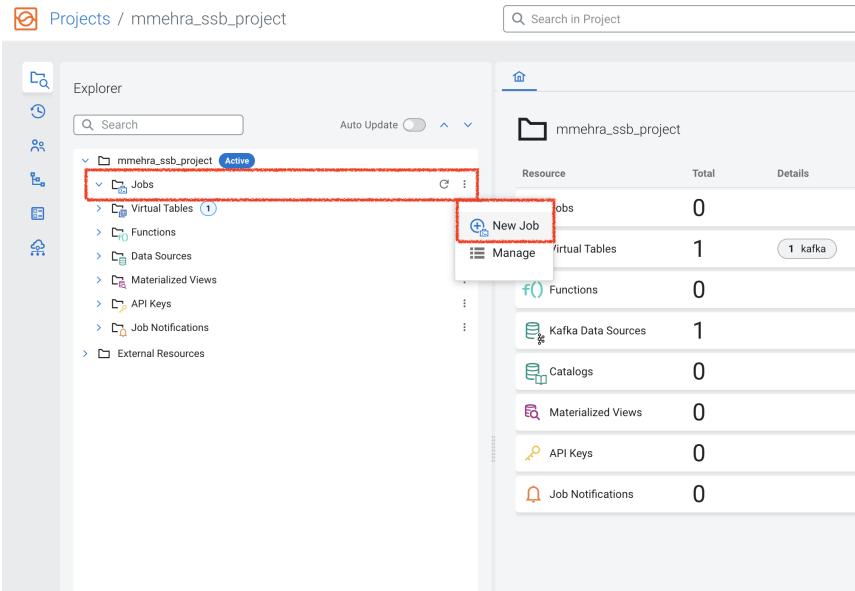
Kafka Data Sources (1)

Catalogs (0)

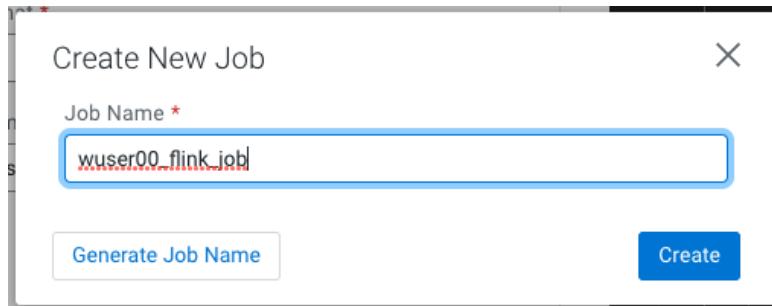
Materialized Views (0)

API Keys (0)

Job Notifications (0)



Give a job name and click **CREATE**



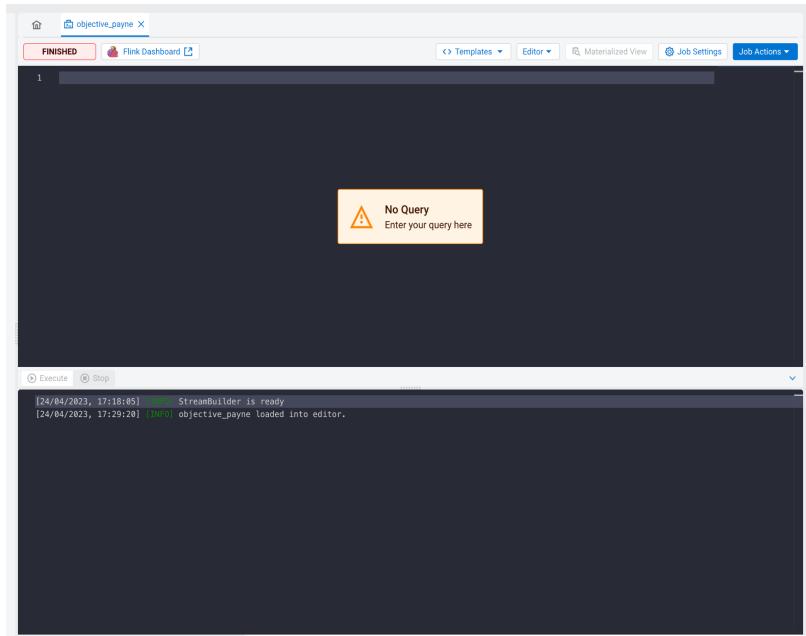
The Query Editor should now show up

FINISHED Flink Dashboard

No Query
Enter your query here

[24/04/2023, 17:18:05] StreamBuilder is ready
[24/04/2023, 17:29:28] objective_payne loaded into editor.

Execute Stop

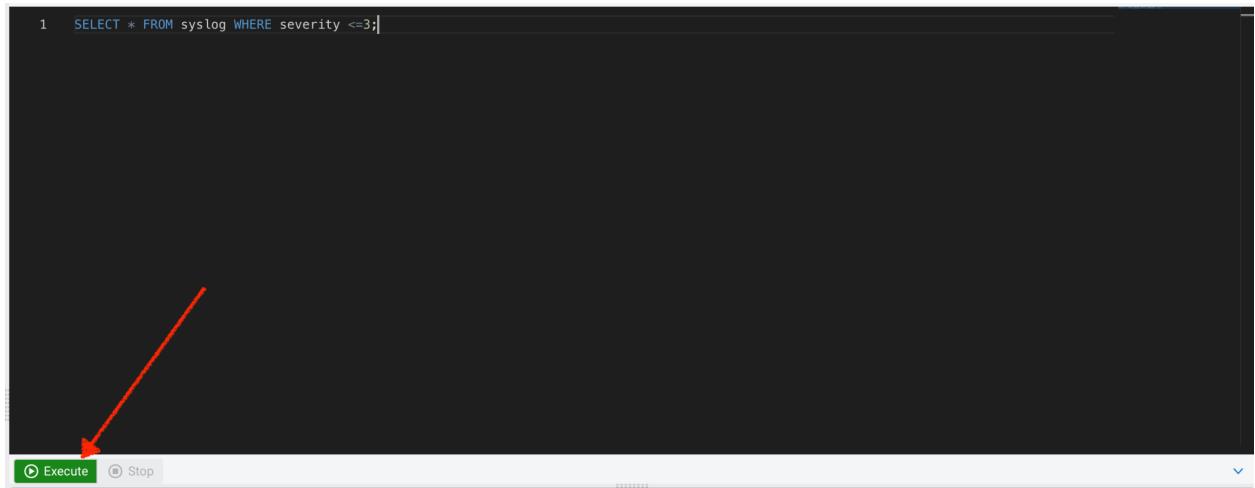


Add the following SQL Statement in the Editor

```
SELECT * FROM {user-id}_syslog_data WHERE severity <=3
```

NOTE : Replace {user-id} with your assigned username

Run the Streaming SQL Job by clicking **Execute**. Also, ensure your {user_id}-syslog-to-kafka flow is running in CDF-PC.



In the Results tab, you should see syslog messages with severity levels <=3

pri...	severity	facility	version	timestamp	hostname	body	appName	procid	messageId	structure...	event_time
114	2	14	1	168335387...	host6.exam...	application...	application1	6063	ID40	{"\$DID": "ev...", "...	2023-05-06...
57	1	7	1	168335392...	host10.exa...	application...	application5	8374	ID20	{"\$DID": "ev...", "...	2023-05-06...
11	3	1	1	168335397...	host7.exam...	application...	application1	7468	ID8	{"\$DID": "ev...", "...	2023-05-06...
179	3	22	1	168335404...	host10.exa...	application...	application10	5088	ID28	{"\$DID": "ev...", "...	2023-05-06...
137	1	17	1	168335413...	host3.exam...	application...	application7	8145	ID40	{"\$DID": "ev...", "...	2023-05-06...
42	2	5	1	168335423...	host4.exam...	application...	application4	702	ID17	{"\$DID": "ev...", "...	2023-05-06...