

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 (wuserxx) 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	

select PARAMETERS	26
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 2: Creation of a Project	77
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.

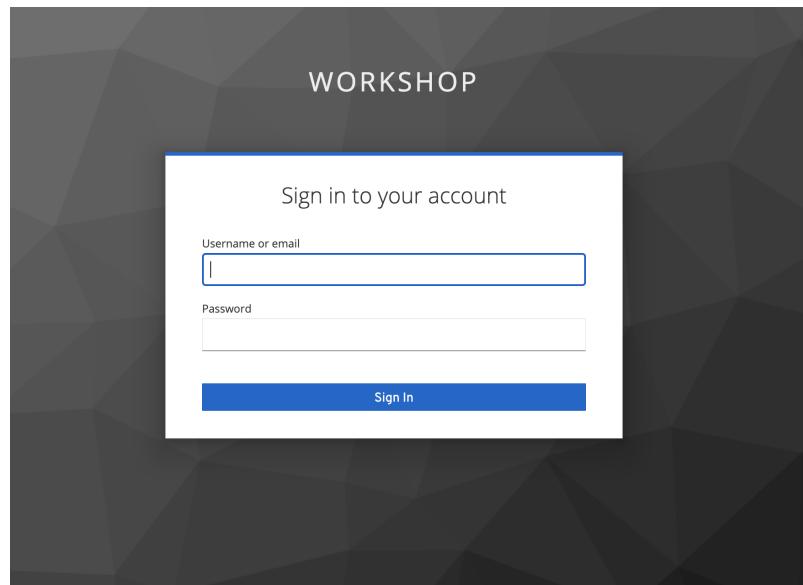
Lab 0 basically talks about verifying different aspects wrt to access and connections before we could begin with the actual steps.

Lab 0 - Introduction and setup

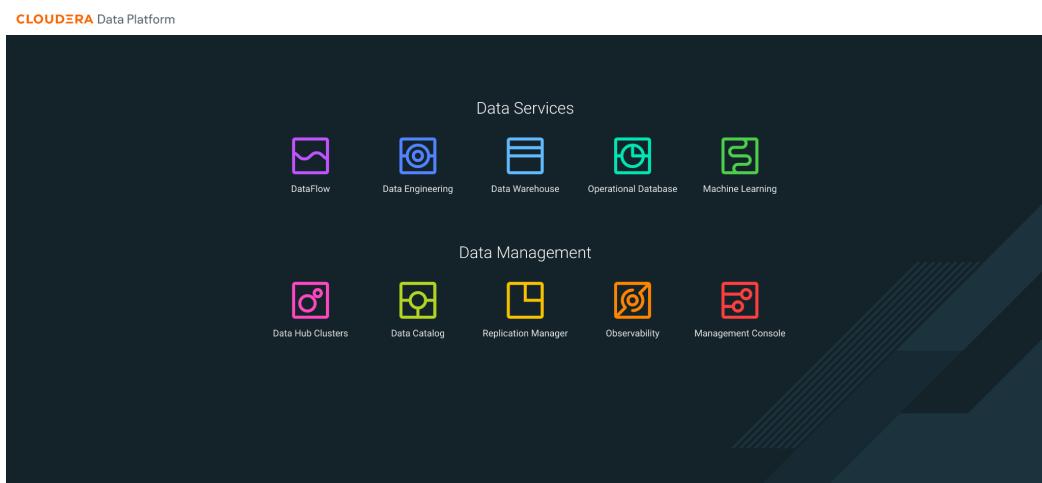
1. Verify access to the workshop environment

- 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.

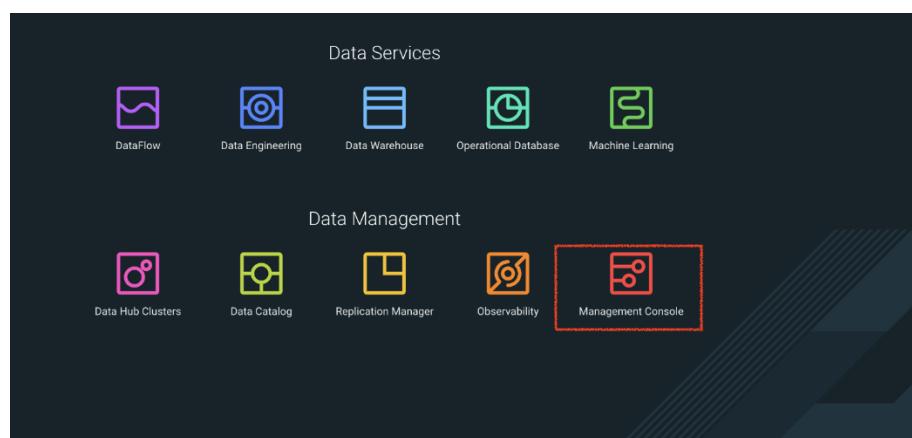


2. Verify permissions in Apache Ranger

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 is a sidebar with navigation links: 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 area displays a table of environments with columns: Status, Name, Cloud Provider, Region, Data Lake, CDP Runtime Version, and Time Created. The table includes rows for environments like 'emeaworkshop-env', 'dss-workshop-env', 'cmi-adb', 'pko-hands-on-workshop-env', 'pse-lsv-env', 'mtna-workshop', 'vrayker-cdp2', and 'pre-workshop'. A search bar at the top of the table allows filtering by environment name.

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

aws pko-hands-on-workshop-env
cm.cdp.environments.us-west-1.d1a4553c-a799-432d-8e54-372cc2ab95f2.environment.e11adbfe-231d-4471-866a-2fb6f3352220
Asia Pacific (Mumbai) - ap-south-1

sdx Data Lake Details

NAME	NODES	SCALE
pko-workshop-dl	2 0 0	Light Duty

STATUS Running **STATUS REASON** Datalake is running **CRN** cm.cdp.datalake.us-west-1.d1a4553c-a799-432d-8e54-372cc2ab95f2:datalake:6a9aed14-c21b-42fe-88c4-e9f7e8a51...

Data Hubs Data Lake FreelPA Cluster Definitions Summary

Ranger Access Manager Audit Security Zone Settings mmehra Last Response Time : 04/20/2023 12:04:09 PM

Service Manager

HDFS cm_hdfs	HBASE cm_hbase	HADOOP SQL Hadoop SQL
YARN cm_yarn flink_wrkshp_cluster_yarn ssb_analytics_cluster_yarn	KNOX cm_knox	SOLR cm_solr
KAFKA cm_kafka flink_wrkshp_cluster_kafka_3306 kafka_smm_cluster_kafka_b3d1 kafka_wrkshp_cluster_kafka_0919 ssb_analytics_cluster_kafka_77ef	NIFI nifi_flow_mgmt_cluster_nifi_NIFI_BASE nifi_wrkshp_cluster_nifi_NIFI_BASE	NIFI-REGISTRY nifi_flow_mgmt_cluster_nifi_registry nifi_wrkshp_cluster_nifi_registry
ATLAS cm_atlas	ADLS cm_adls	KUDU cm_kudu
OZONE cm_ozone	SCHEMA-REGISTRY kafka_smm_cluster_schemaregistry kafka_wrkshp_cluster_schemaregistry	KAFKA-CONNECT cm_kafka_connect kafka_smm_cluster_kafka_connect

2.2 Kafka Permissions

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

KAFKA

cm_kafka	
flink_wrkshp_cluster_kafka_3306	
kafka_smm_cluster_kafka_b3d1	
ssb_analytics_cluster_kafka_77ef	

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_3e7187a6	cruisecontrol ssb streamsmgr kafka kafka_mirror_maker streamsrepmgr rangerlookup - Less..	
105	all - topic	--	Enabled	Enabled	--	c_ranger_admins_3e7187a6	cruisecontrol ssb streamsmgr kafka kafka_mirror_maker streamsrepmgr rangerlookup - Less..	

- All-consumergroup

Allow Conditions:

- all-topic

Policy Details:

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_6276836c]	[sub streammsgmgr kafka schemaregistry + More...]	[Edit Delete]
160	all - serde	--	Enabled	Enabled	--	[c_ranger_admins_6276836c]	[sub streammsgmgr kafka schemaregistry + More...]	[Edit Delete]
163	all - schema-group, schema-metadata	--	Enabled	Enabled	--	[c_ranger_admins_6276836c]	[sub streammsgmgr kafka schemaregistry + More...]	[Edit Delete]
165	all - schema-group, schema-metadata, sch...	--	Enabled	Enabled	--	[c_ranger_admins_6276836c]	[sub streammsgmgr kafka schemaregistry + More...]	[Edit Delete]
167	all - registry-service	--	Enabled	Enabled	--	[c_ranger_admins_6276836c]	[sub streammsgmgr kafka schemaregistry + More...]	[Edit Delete]
169	all - schema-group, schema-metadata, sch...	--	Enabled	Enabled	--	[c_ranger_admins_6276836c]	[sub streammsgmgr kafka schemaregistry + More...]	[Edit Delete]

Policy Details:

Policy Type	Access		
Policy ID	101		
Policy Name *	all - schema-group, schema-metadata, schema-t	Enabled <input checked="" type="radio"/>	Normal <input type="radio"/>
Policy Label	Policy Label		
schema-grn	<input type="text"/> * <input type="button" value="X"/>	<input type="button" value="Include"/>	
Schema Name	<input type="text"/> * <input type="button" value="X"/>	<input type="button" value="Include"/>	
schema-brn	<input type="text"/> * <input type="button" value="X"/>	<input type="button" value="Include"/>	
schema-ver	<input type="text"/> * <input type="button" value="X"/>	<input type="button" value="Include"/>	
Description	Policy for all - schema-group, schema-metadata, schema-branch, schema-version		
Audit Logging	<input checked="" type="radio"/> Yes		

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: wuser00@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 various navigation options like Dashboard, Environments, Data Lakes, and User Management. The main area is titled 'Users : wuser00@workshop.com'. It displays detailed information for the user 'wuser00@workshop.com', including Name, Email, Workload User Name, CHN, Tenant ID, Identity Provider, Last Interactive Login, Profile Management, and Workload Password. The 'Workload Password' field is highlighted with a red box. Below this, there are tabs for Access Keys, Roles, Resources, Groups, and SSH Keys. A message at the bottom states 'No access keys found' with a '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 two input fields: 'Password' and 'Confirm Password', both with red asterisks indicating they are required. Below these 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.

- Click the button **Set Workload Password**.

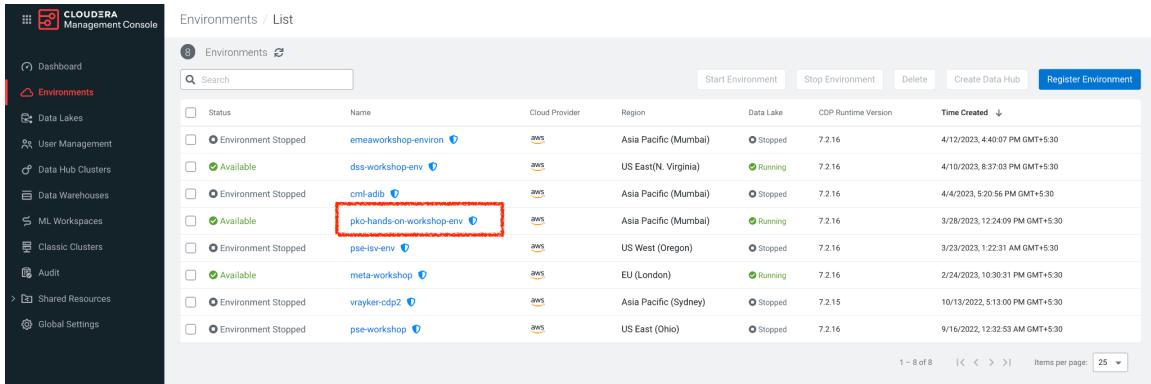
4. Obtain the Kafka Broker List

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

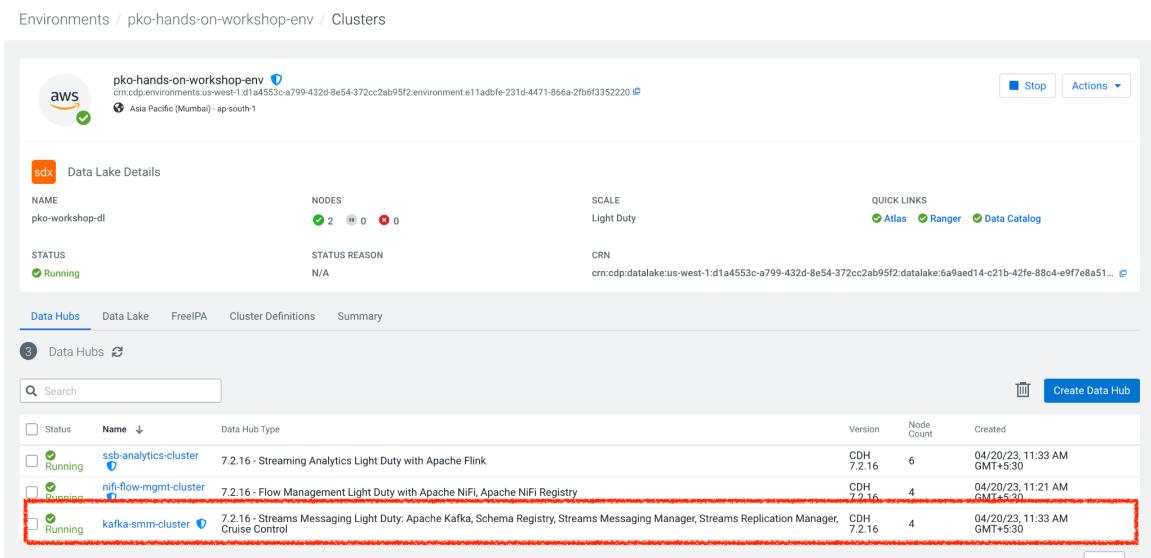


The screenshot shows the Cloudera Management Console interface. On the left, there's a sidebar with various navigation options like Dashboard, Environments, Data Lakes, User Management, etc. The main area is titled 'Environments / List'. It contains a table with columns: Status, Name, Cloud Provider, Region, Data Lake, CDP Runtime Version, and Time Created. There are eight environments listed:

Status	Name	Cloud Provider	Region	Data Lake	CDP Runtime Version	Time Created
Environment Stopped	emeaworkshop-environ	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-lsv-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	vayer-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

At the bottom right of the table, there are pagination controls (1 – 8 of 8) and a dropdown for 'Items per page' set to 25.

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



The screenshot shows the Data Hub Details page for the 'pko-hands-on-workshop-env' environment. At the top, there's a summary card for the Data Hub. Below it, the 'Data Hubs' tab is selected, showing a table of Data Hubs:

Name	Nodes	Scale	Quick Links
pko-workshop-dl	2 nodes (0 active, 0 failed)	Light Duty	Atlas, Ranger, Data Catalog
STATUS	Running	CRN	crn:cdf:datalake:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:datalake:6a9aed14-c21b-42fe-88c4-e9f7e8a51...

Below this table, there are tabs for Data Hubs, Data Lake, FreeIPA, Cluster Definitions, and Summary. The Data Hubs tab is currently active. At the bottom, there's a search bar and a 'Create Data Hub' button.

A red box highlights the 'kafka-smm-cluster' row in the Data Hubs table, which is described as '7.2.16 - Streams Messaging Light Duty: Apache Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control'.

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. At the top, it displays cluster statistics: STATUS (Running), NODES (4), CREATED AT (04/20/23, 11:33 AM GMT+5:30), and CLUSTER TEMPLATE (7.2.16 - Streams Messaging Light Duty: Apache Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control). Below this, the STATUS REASON indicates "Cluster started". Under Environment Details, it shows NAME (pko-hands-on-workshop-env), DATA LAKE (pko-workshop-dl), CREDENTIAL (pko-hands-on-workshop-cred), REGION (ap-south-1), and AVAILABILITY ZONE (N/A). Services listed include CM-UI, Schema Registry, Streams Messaging Manager (highlighted with a red box), Token Integration, and Cloudera Manager Info. The CM URL is https://kafka-smm-cluster-gateway.pko-hand.dp5i-5vkq.cloudera.site/kafka-smm-cluster/cdp-proxy/cm/home/. The Event History tab is selected, followed by Autoscale, Endpoints (5), Tags (4), Nodes, Network, Load Balancers, Telemetry, Repository Details, Image Details, Recipes (0), Cloud Storage, Database, and Upgrade. At the bottom, there are filters for Events, Show All, Autoscale, Cluster, and a DOWNLOAD button.

Step 3 : Select Brokers from the left tab

The screenshot shows the Kafka Streams Messaging Manager interface with the Brokers tab selected. A red box highlights the "Brokers" section. The main table displays 14 Producers, 3 Brokers, 33 Topics, and 3 Consumer Groups. The Brokers table includes columns for NAME, DATA IN, DATA OUT, MESSAGES IN, CONSUMER GROUPS, and CURRENT LOG SIZE. Consumer Groups are also listed on the right side of the screen.

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

Step 4 : Save the broker list

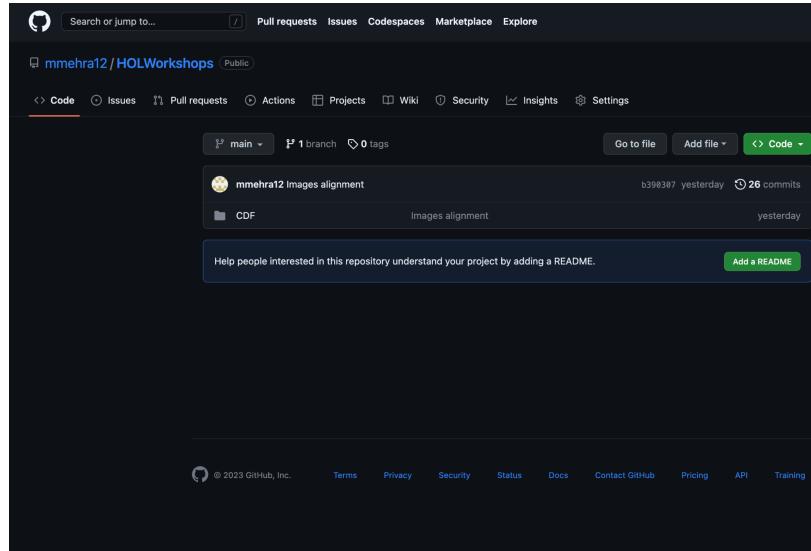
Brokers								Cluster: kafka-b3d1
Total Bytes In 680 MB	Total Bytes Out 2 MB	Produced Per Sec 307	Fetched Per Sec 4,928	Active Controllers 1	Unclean Elections 0	Request Pool Usage 0.00%	Search	🕒 30 minutes
Brokers (3)								
NAME	THROUGHPUT	MESSAGES IN	PARTITIONS	REPLICAS	LOG SIZE	REMAINING STORAGE		
1546335432 kafka-smm-cluster-corebroker1.pko-hand.dp5i-5vkq.cloudera.site:9093	682 MB	2.3m	94	261	2 GB	982 GB		
1546335453 kafka-smm-cluster-corebroker0.pko-hand.dp5i-5vkq.cloudera.site:9093	238 KB	3.4k	100	270	6 GB	978 GB		
1546335411 kafka-smm-cluster-corebroker2.pko-hand.dp5i-5vkq.cloudera.site:9093	500 KB	6.5k	98	263	4 GB	980 GB		

Example :

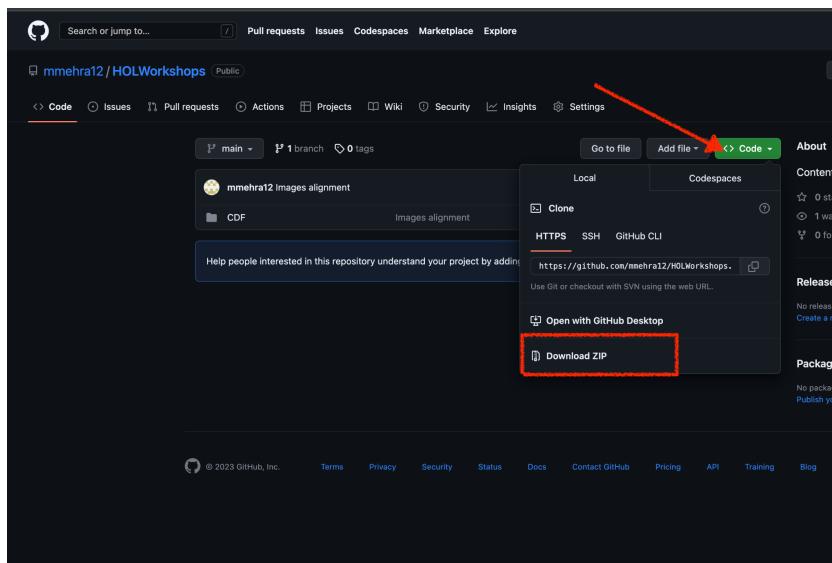
kafka-smm-cluster-corebroker1.pko-hand.dp5i-5vkq.cloudera.site:9093
kafka-smm-cluster-corebroker0.pko-hand.dp5i-5vkq.cloudera.site:9093
kafka-smm-cluster-corebroker2.pko-hand.dp5i-5vkq.cloudera.site:9093

5. Download Resources from GitHub

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

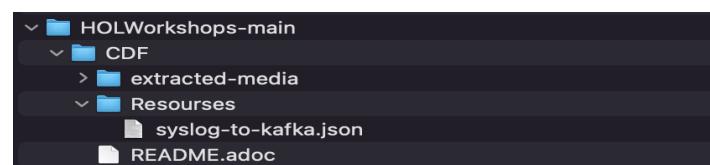


Step 2 : Download the repo as a zip file



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

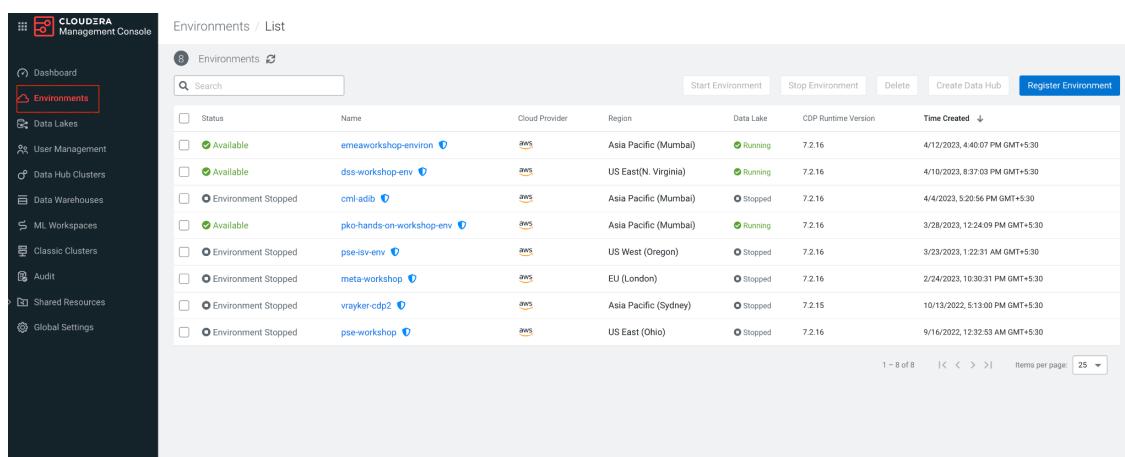
6. Unlock your KeyTab

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

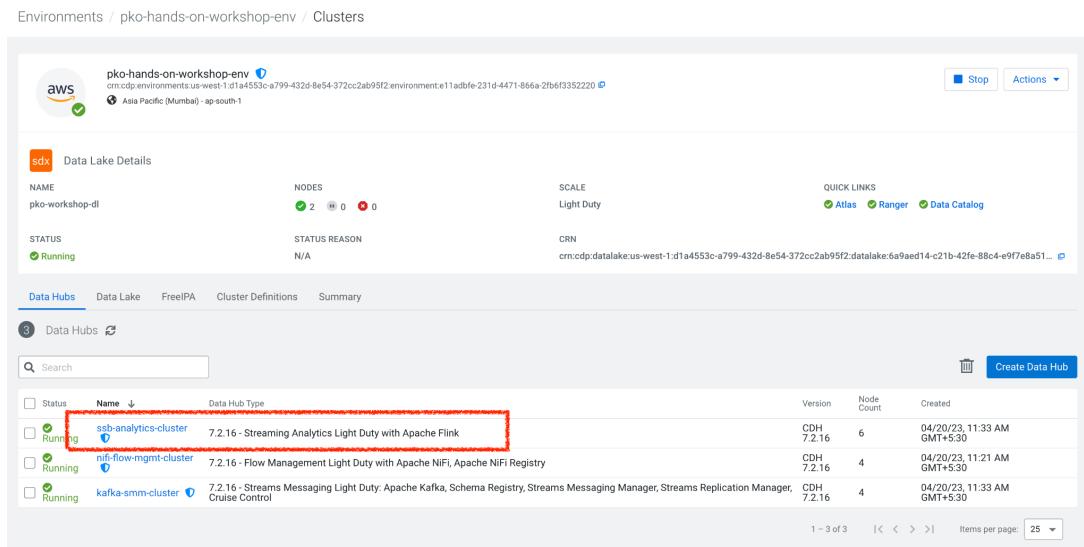
Step 1 : Go to the SSB Data Hub

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	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-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
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)



Name	Nodes	Scale	Quick Links
pko-workshop-dl	2	Light Duty	Atlas Ranger Data Catalog
STATUS	STATUS REASON	CRN	
Running	N/A	cm.cdp:datalake:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:datalake:6a9aed14-c21b-42fe-88c4-e9f7e8a51...	

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

Data Hubs / ssb-analytics-cluster / Event History

ssb-analytics-cluster

Actions

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	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.dpSi-5vkq.cloudera.site/ssb-analytics-cluster/cdp/proxy/cmft/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

Projects

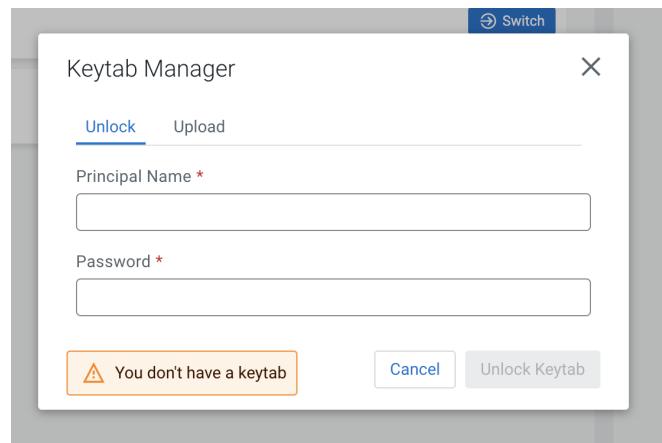
My Projects

Search Reload Import New Project

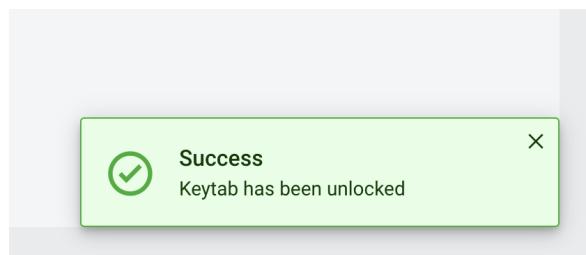
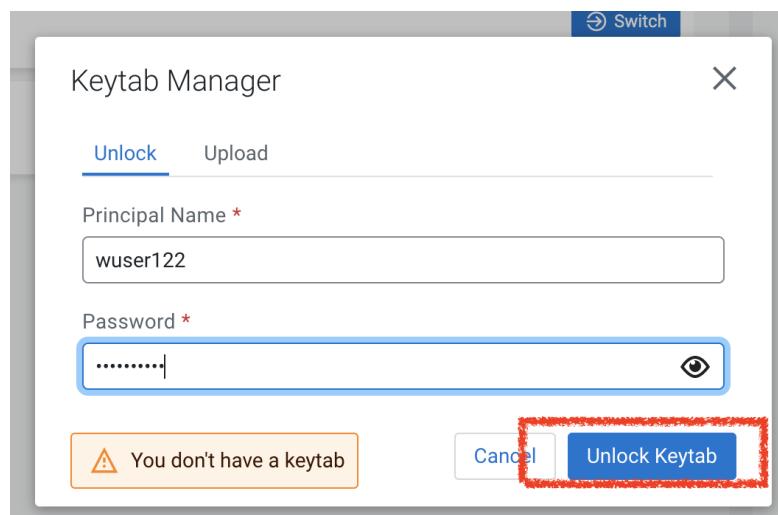
Name	ID	MV Prefix	Actions
wuser122_hol_workshop Active	65478db5		Open Delete
wuser122_default	3756dd72		Switch
ssb_default	ffffffff		Switch

wuser122 Manage Keytab

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



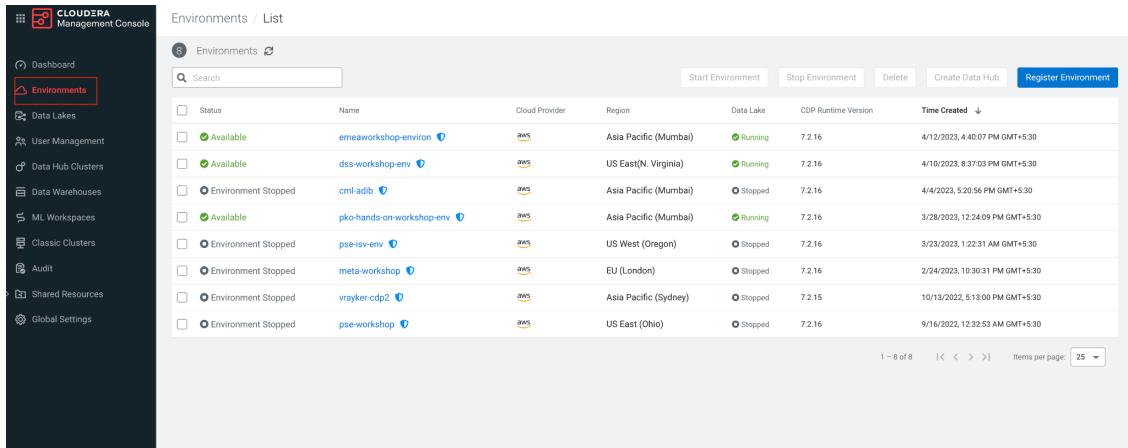
Step 5 : Click on unlock KeyTab



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

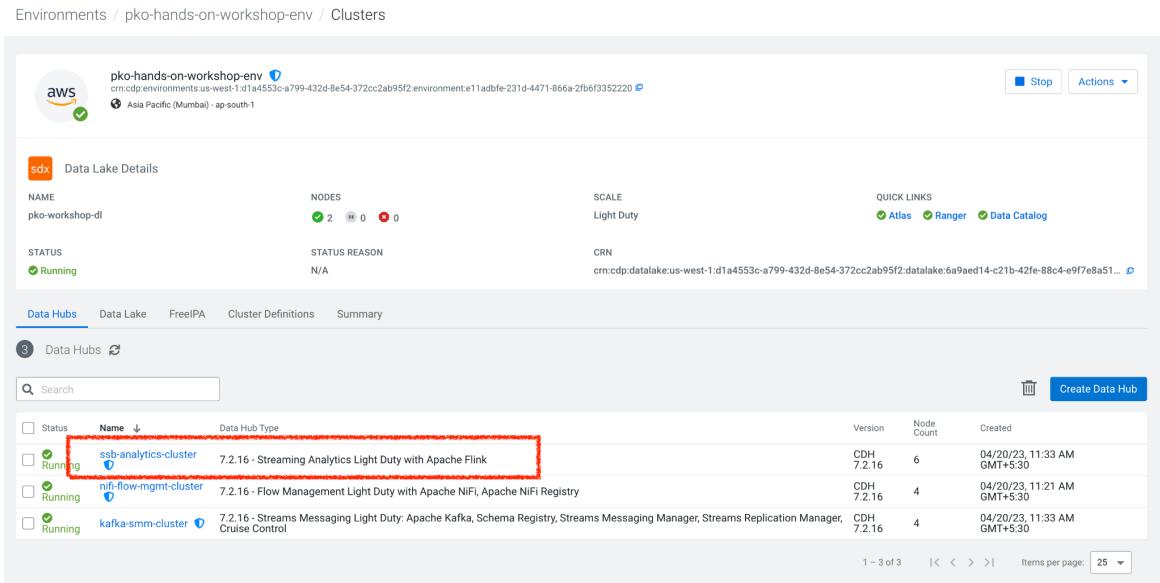


The screenshot shows the CloudERA Management Console interface. On the left, there's a sidebar with various navigation options like Dashboard, Environments (which is selected and highlighted in red), Data Lakes, User Management, Data Hub Clusters, ML Workspaces, Classic Clusters, Audit, Shared Resources, and Global Settings. The main area is titled 'Environments / List' and contains a table with the following data:

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-adib	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

At the bottom right of the table, there are pagination controls (1 - 8 of 8) and a dropdown for 'Items per page' set to 25.

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



The screenshot shows the 'Data Hub Details' page for the environment 'pko-hands-on-workshop-env'. At the top, there's a summary card for the environment, followed by tabs for Data Hubs, Data Lake, FreelPA, Cluster Definitions, and Summary. The 'Data Hubs' tab is currently selected. Below the tabs, there's a search bar and a table listing running 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

At the bottom right of the table, there are pagination controls (1 - 3 of 3) and a dropdown for 'Items per page' set to 25.

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	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.dpSi-5vkq.cloudera.site/ssb-analytics-cluster/cdp/proxy/cm/home/	7.9.0	7.2.16-1.cdf7.2.16.p2.38683602	Command logs , Service logs

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

Projects

My Projects

Name	ID	MV Prefix
wuser122_hol_workshop	65478db5	
wuser122_default	3756d572	
ssb_default	!!!!!!!	

wuser122

Manage Keytab

Keytab Manager

Lock Upload

Principal Name *

You have a keytab

Cancel Lock Keytab

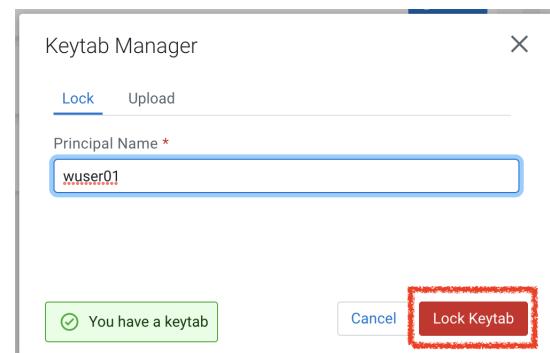
If you get the following dialog box it means that your Keytab is already unlocked. **But it would be necessary to reset here by locking it and unlocking it again using your newly set workload password**

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

Example : wuserXY

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.

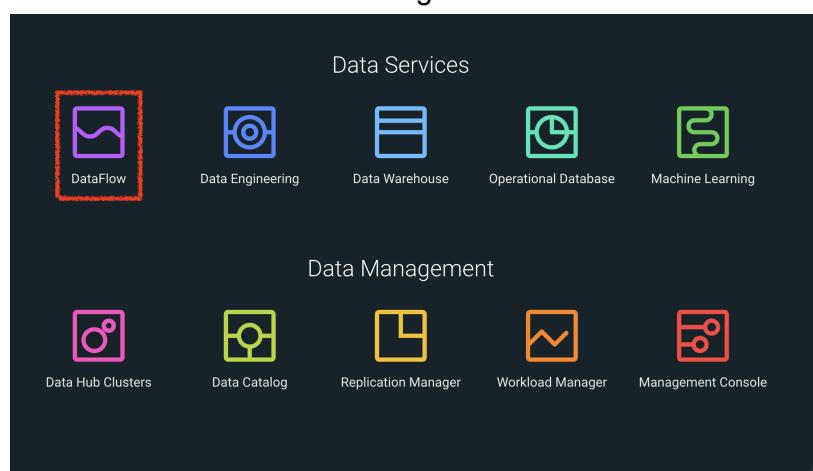
The following is a step by step guide in building a data flow for use within CDF-PC.

2. Building the Data Flow

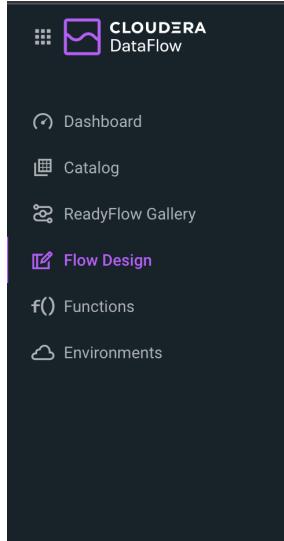
2.1. Create the canvas to design your flow

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)

A screenshot of the Flow Design interface. The top navigation bar shows 'Flow Design'. Below it, the 'All Drafts' section has a search bar and a red box highlights the 'Create Draft' button. A status message at the bottom right says 'REFRESHED: 7 seconds ago'.

Step 4: Select the appropriate environment

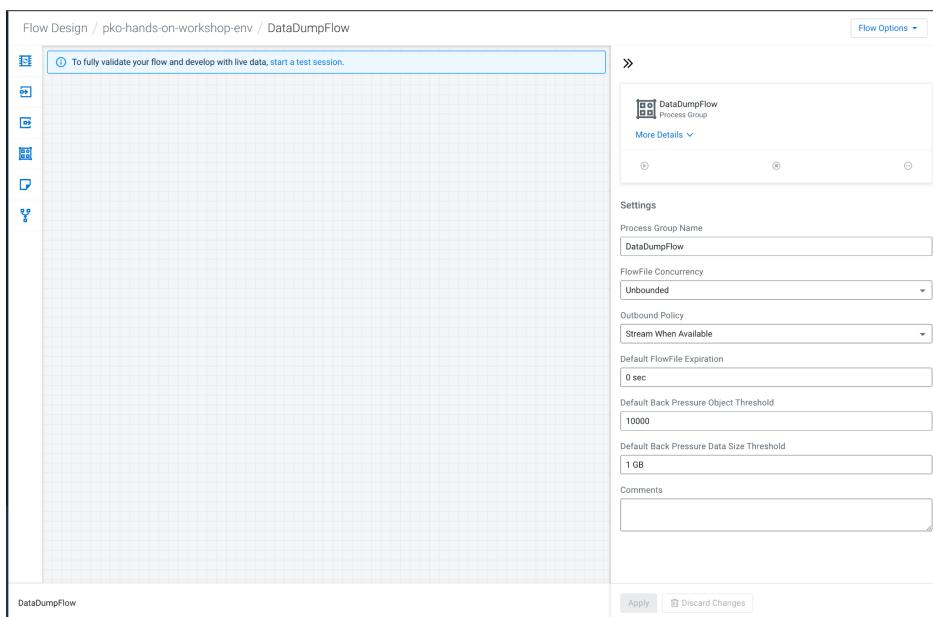
Select the appropriate environment as part of the workspace and give your flow a name and click on **CREATE**

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

Draft Name : {user_id}_datadump_flow
Example : wuserXY_datadump_flow

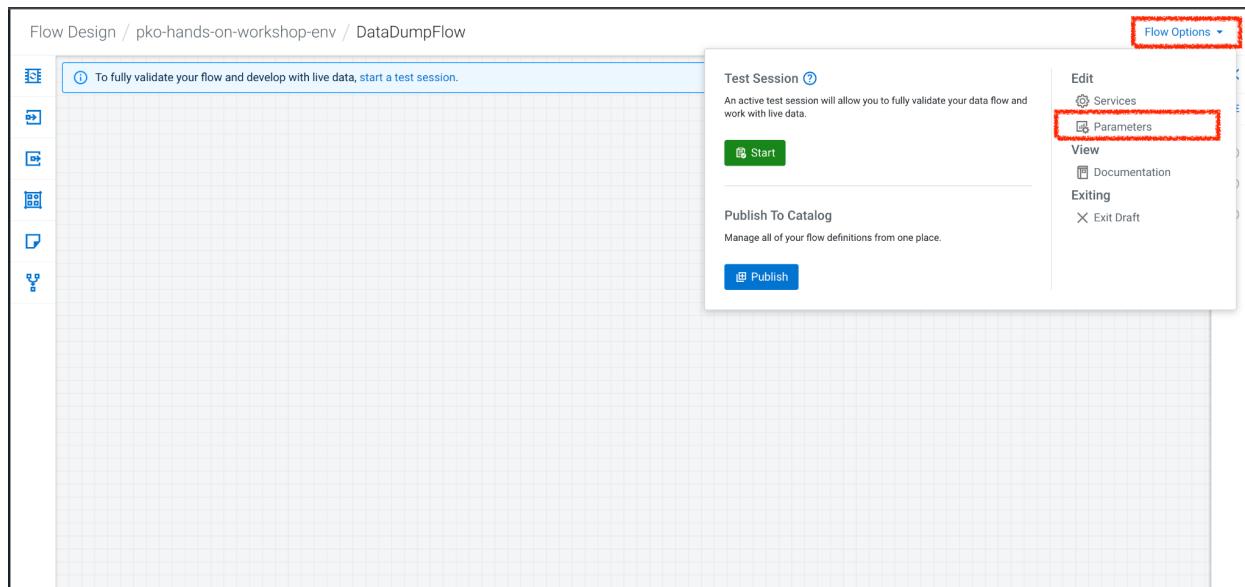
A screenshot of the 'Create New Draft' dialog box. It asks to 'Select the target workspace' and shows a dropdown menu with 'aws pko-hands-on-workshop-env' selected. The 'Draft Name' field contains 'wuser00_datadump_flow' and has a validation message 'Draft name is valid'. At the bottom are 'Cancel' and 'Create' buttons.

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



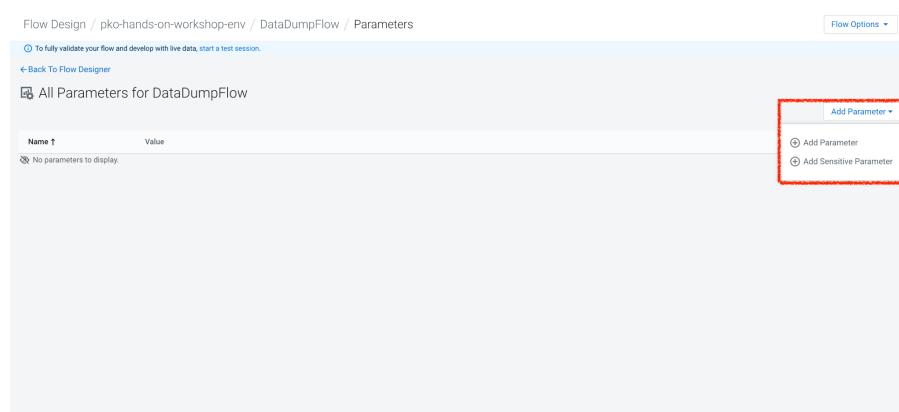
2.2. Adding new parameters

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



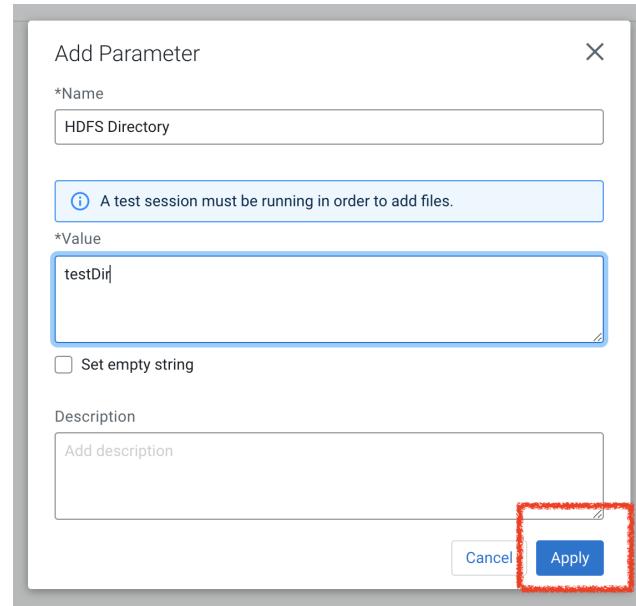
Step 2: Configure Parameters

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**.

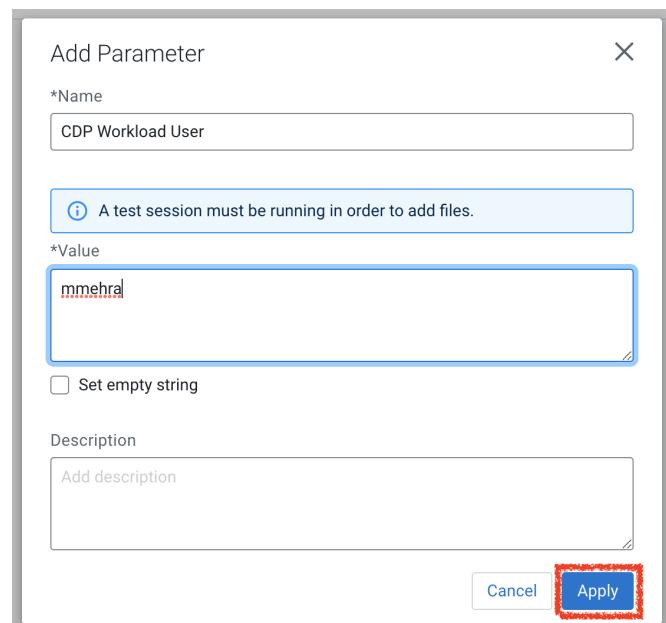


We need to add the following parameters.

- 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 : wuser01
 - IMPORTANT: do not add the domain '@workload.com'



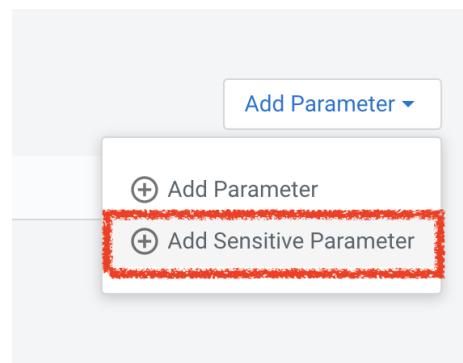
- CDP Workload User Password - [Sensitive Field]

- Selection under Add Parameter : **Add**

Sensitive Parameter

- Name : CDP Workload User Password
- Value : <Workload Password set by yourself in Lab 0>

■ EXAMPLE : Wuser@2021



Add Sensitive Parameter

*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**

Click **APPLY CHANGES**

All Parameters for hostmm_datadump_flow		
Name	Value	Changed
CDP Workload User	wuser00	<input checked="" type="checkbox"/> Modified
CDP Workload User Password <small>Sensitive value set</small>	<input checked="" type="checkbox"/> Modified
S3 Directory	LabData	<input checked="" type="checkbox"/> Modified

Apply Changes **Discard Changes**

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

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

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:

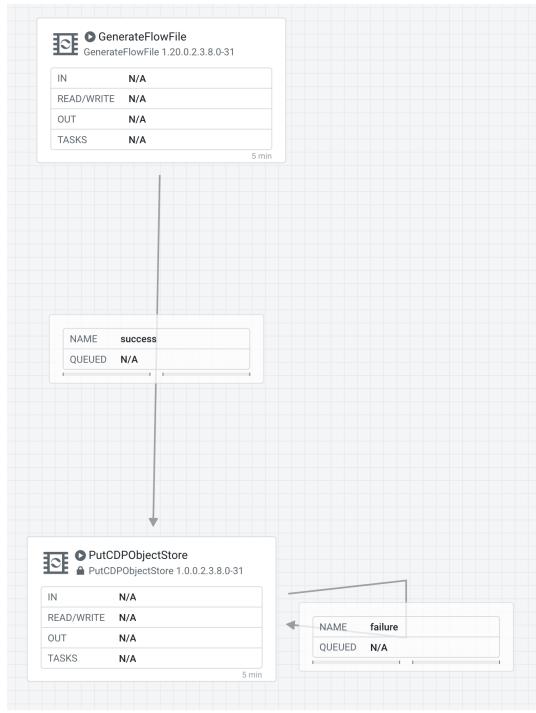
1. Open a processor's configuration and proceed to the properties tab.
2. Enter: #{{
3. Hit ‘control+spacebar’

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

2.3. Create the Flow

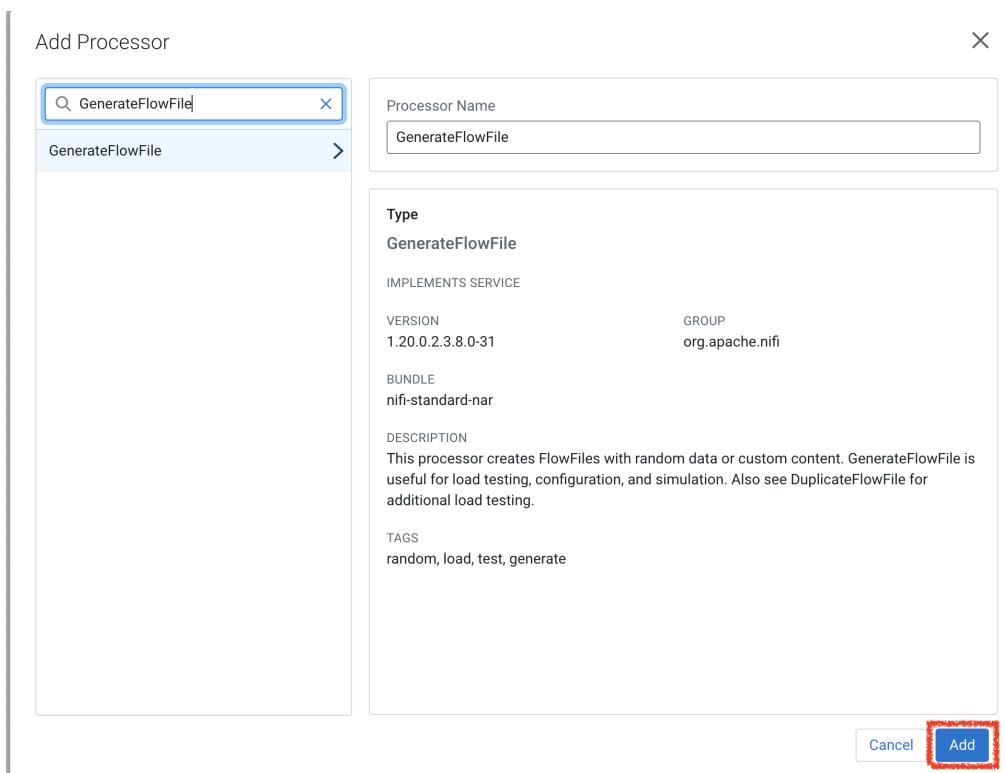
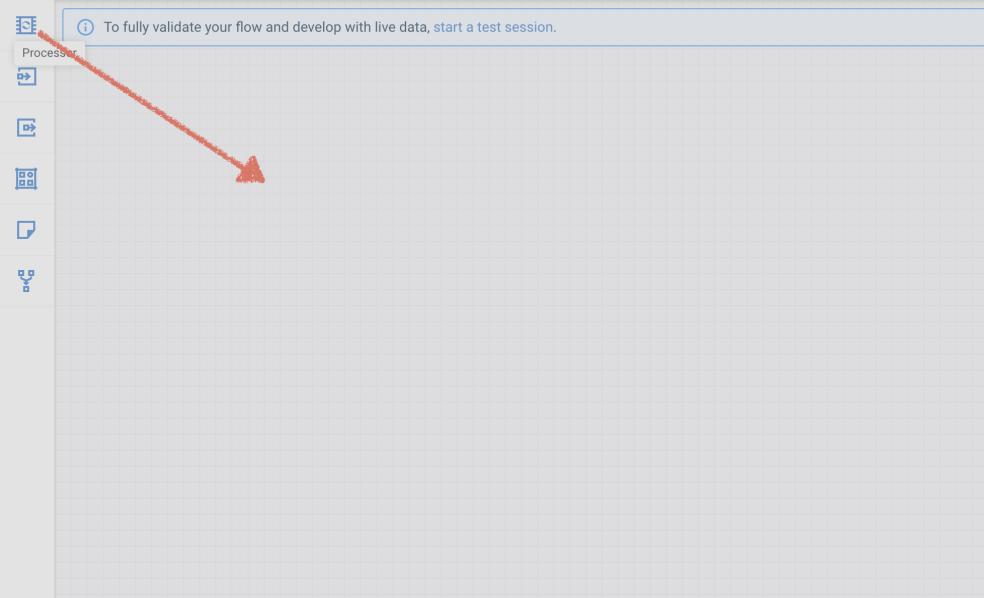
Let's go back to the canvas to start designing our flow. This flow will contain 2 Processors:

- **GenerateFlowFile** - Generates random data
- **PutCDPObjectStore** - Loads data into HDFS(S3)
- Our final flow will look like this:



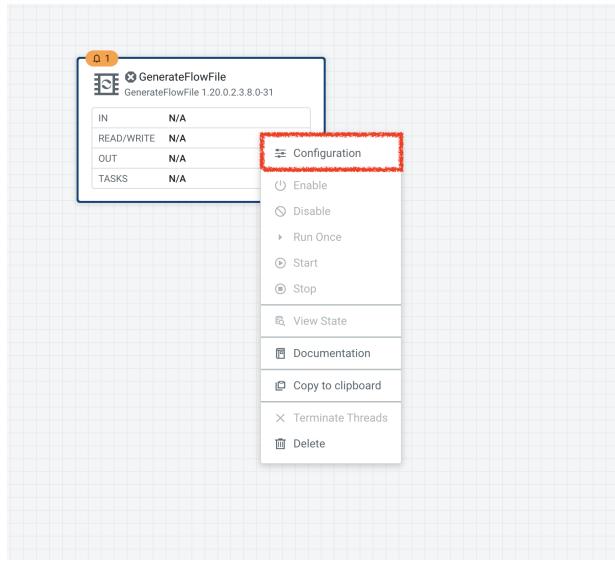
Step 1: Add **GenerateFlowFile** processor

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

DataGenerator

*Penalty Duration ⓘ

30 sec

*Yield Duration ⓘ

1 sec

*Bulletin Level ⓘ

WARN

Comments

Scheduling

*Scheduling Strategy ⓘ

Timer Driven

*Concurrent Tasks ⓘ

1

*Run Duration ⓘ

0ms

*Run Schedule ⓘ

30 sec

*Execution ⓘ

All Nodes

Properties

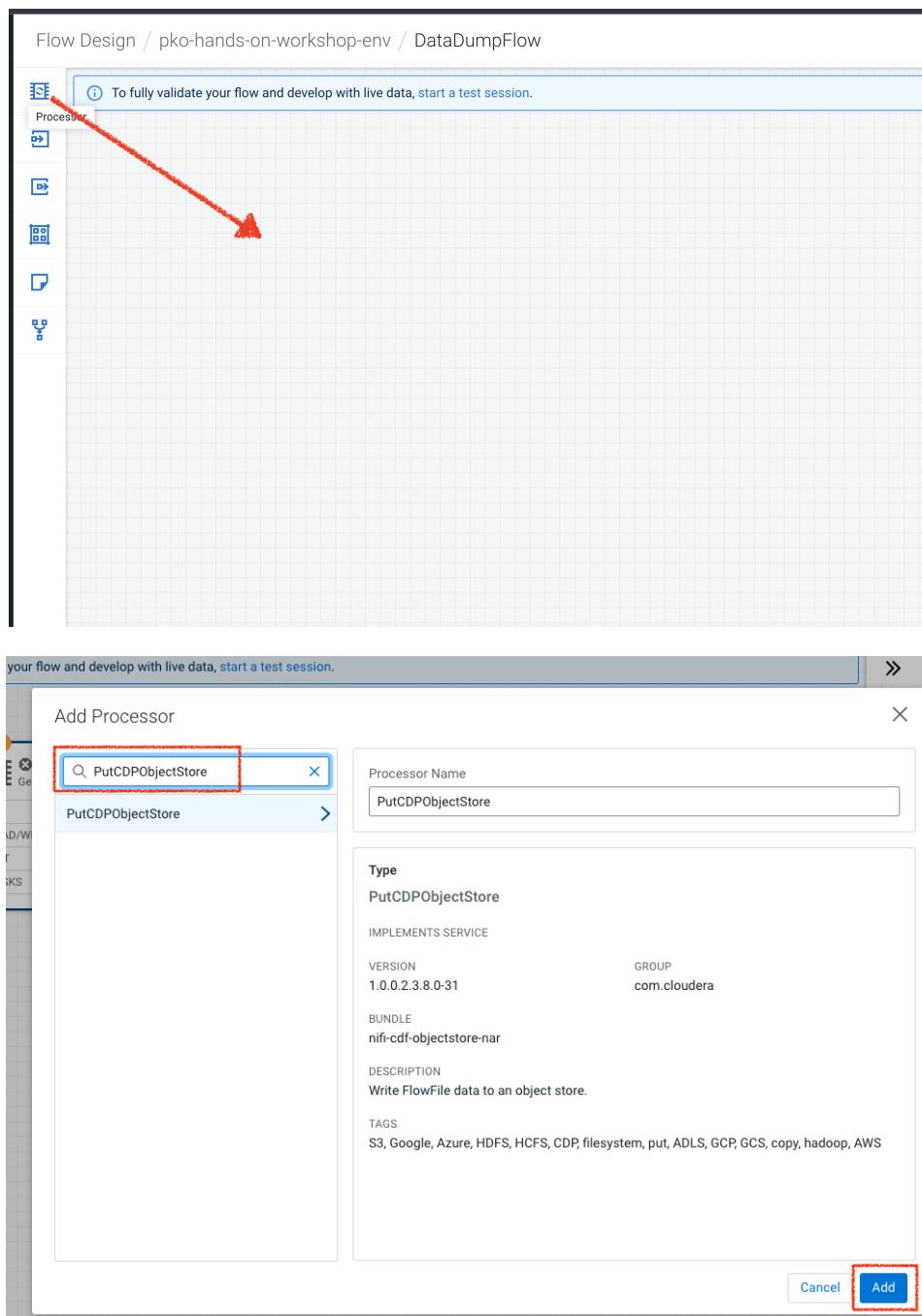
[Add Property](#)

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”

 PutCDPObjectStore
PutCDPObjectStore 1.0.0.2.3.8.0-31

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

*Number of Retry Attempts ②

10

Click APPLY

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

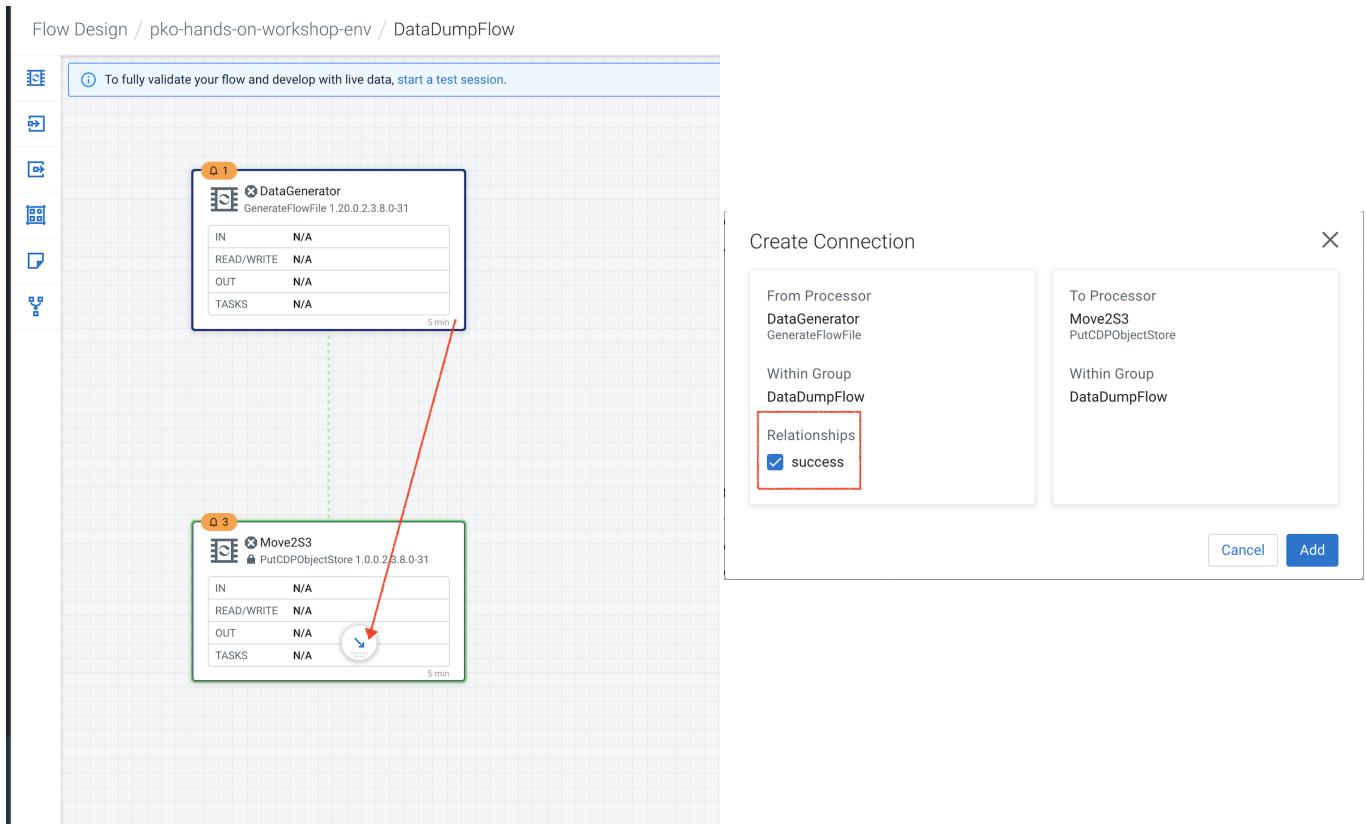
10 mins

Apply

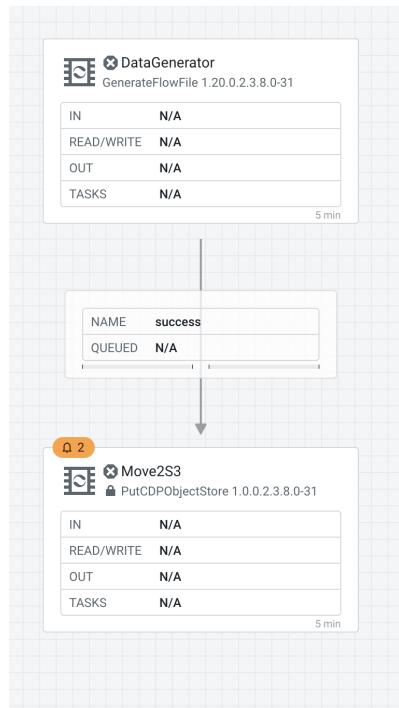
Discard Changes

Step 5: Create connection between processors

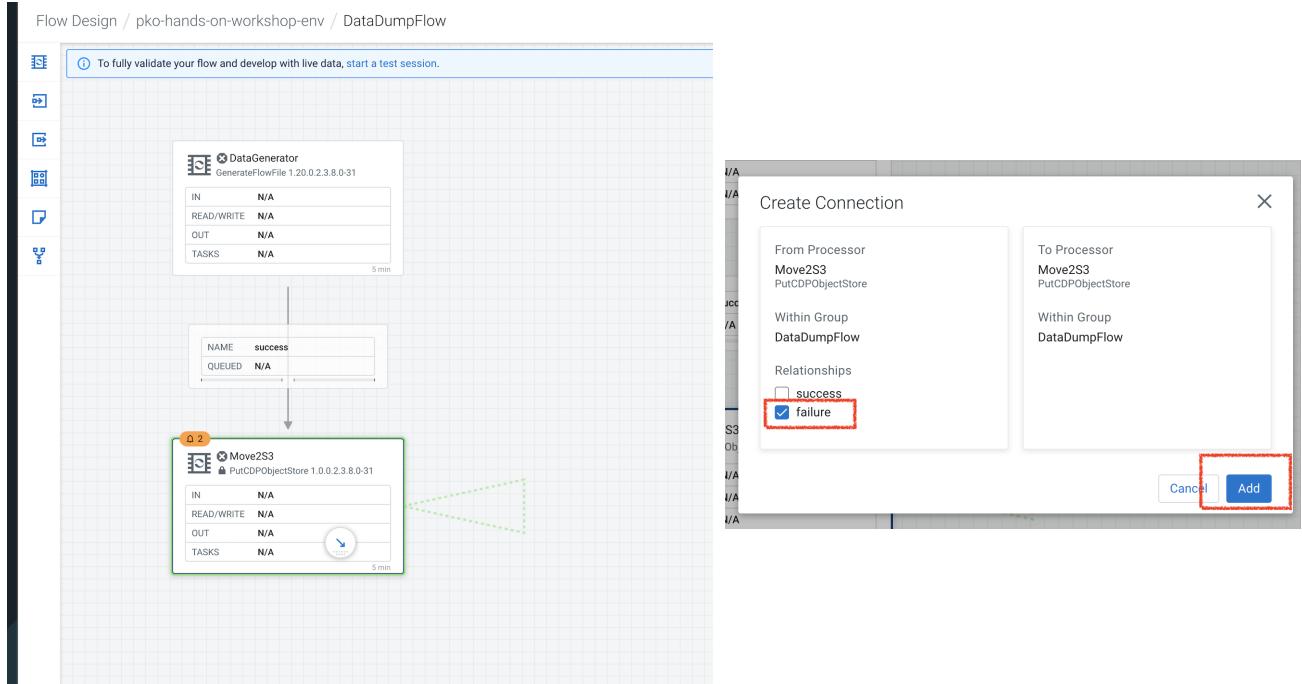
Connect the two processors by dragging the arrow from **DataGenerator** processor to the **Move2S3** processor and select on **SUCCESS** relation and click **ADD**



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:

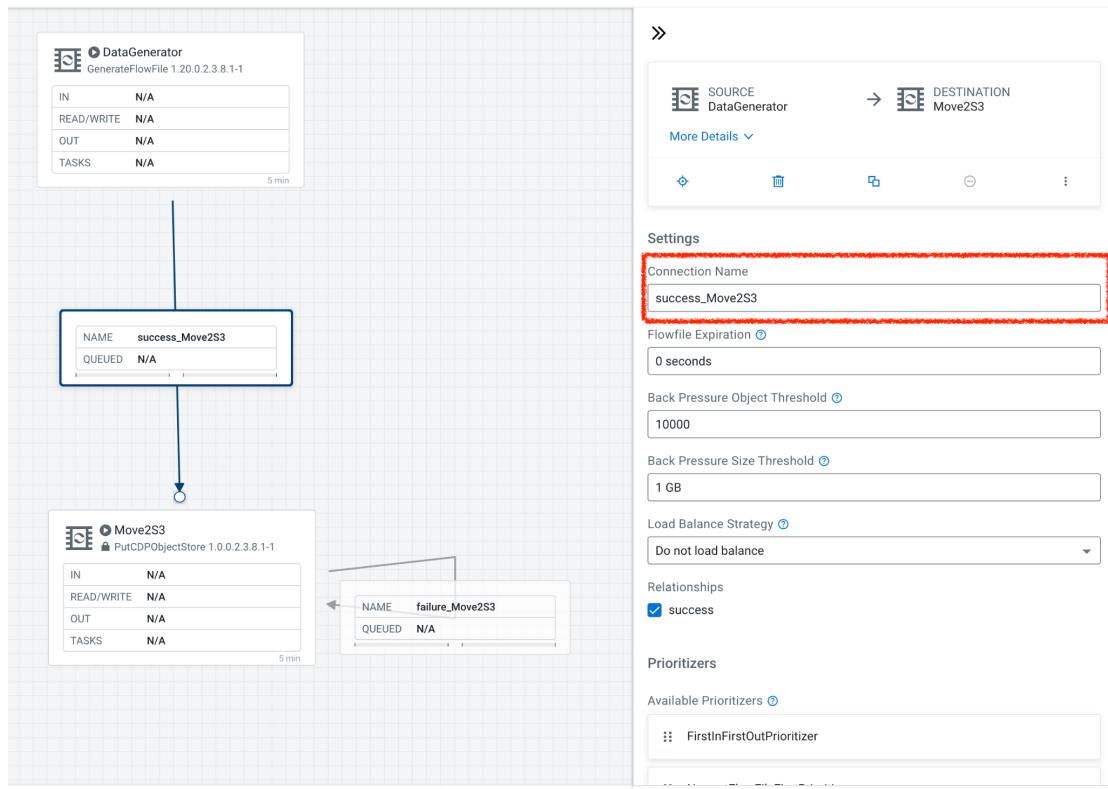


2.4. Naming the queues

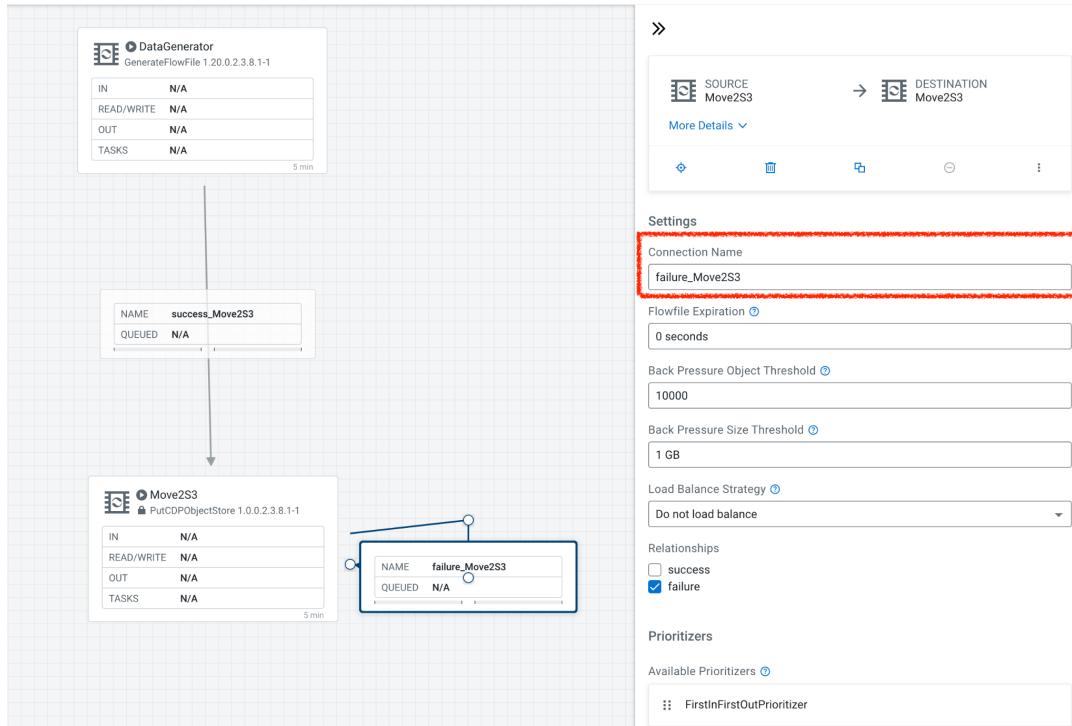
Providing unique names to all queues is very important as they are used to define Key Performance Indicators upon which CDF-PC will auto-scale.

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**.

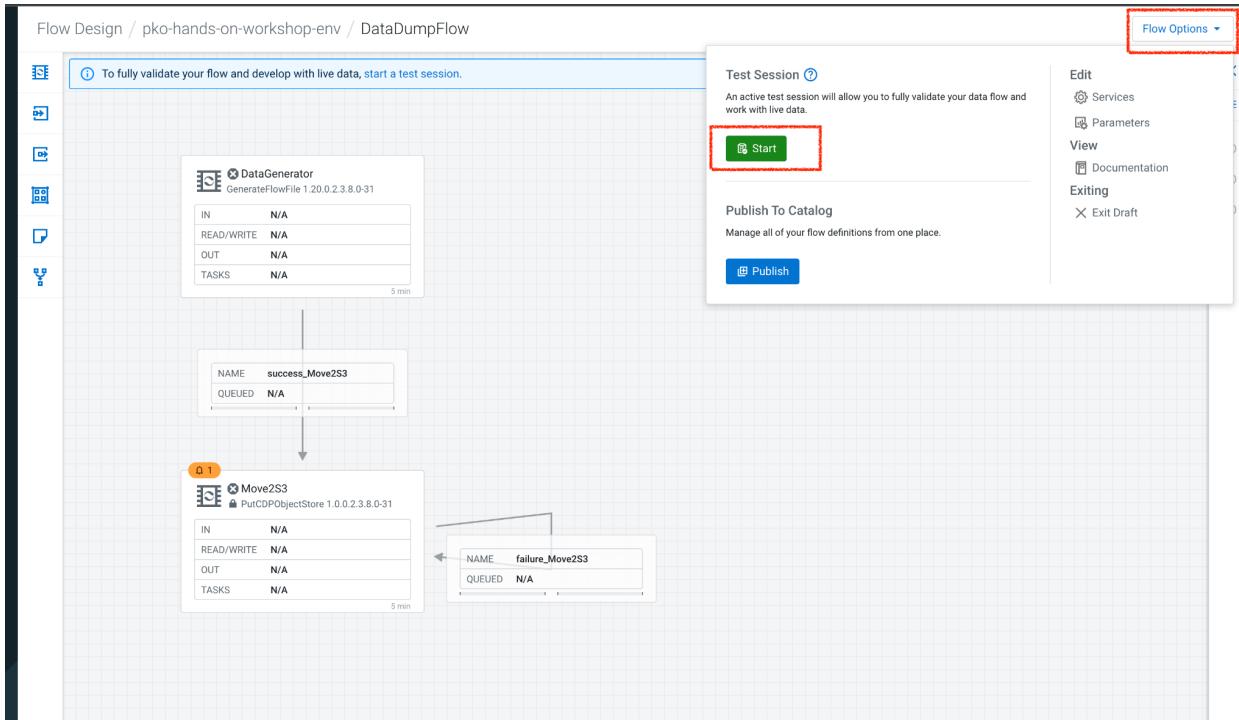


3. Testing the Data Flow

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**

NiFi Configuration

NiFi Runtime 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.

Inbound Connections

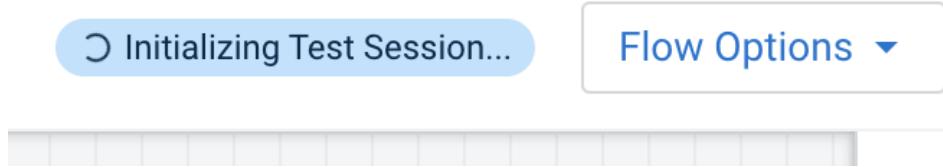
Allow NiFi to receive data

Custom NAR Configuration

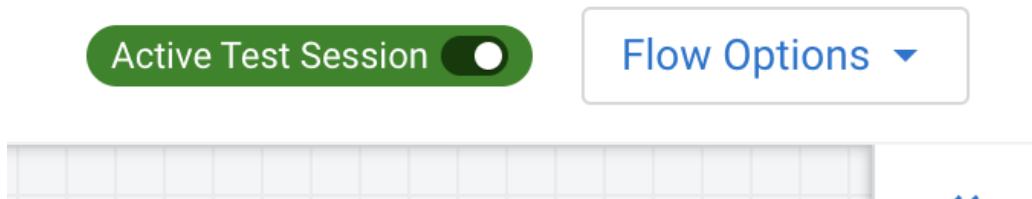
This flow deployment uses custom NARs

Start 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

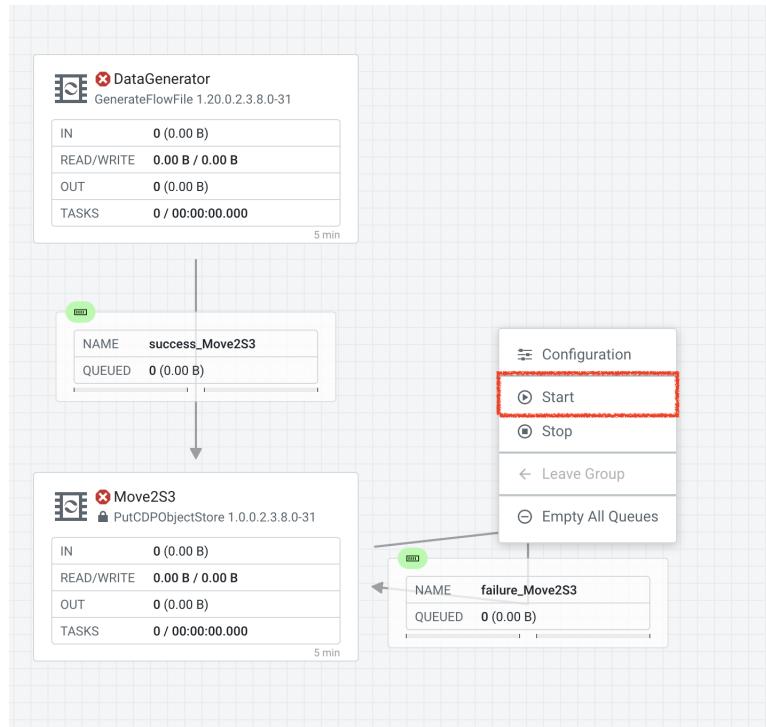


Once the Test Session is ready you will see the following message on the top right corner of your screen.

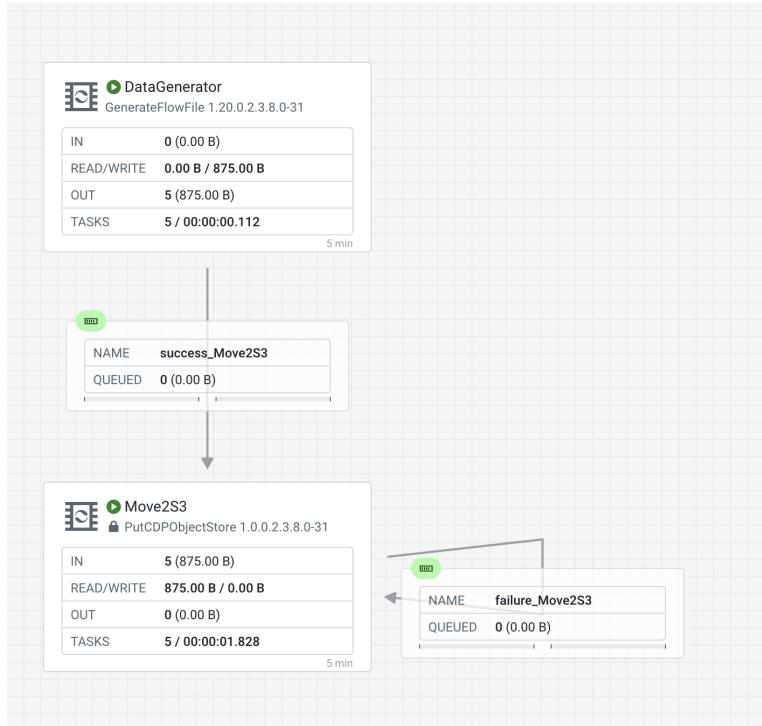


Step 2: Run the flow

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/

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](#)

[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-ac99-0c6127968dff	-	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
8bf998a-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
cba5becb-6d4e-430c-9dfc-c477d992e30c	-	March 29, 2023, 20:51:24 (UTC+05:30)	175.0 B	Standard
cab44364-b35f-4cad-a0f8-b2b885873eb	-	March 29, 2023, 20:51:57 (UTC+05:30)	175.0 B	Standard

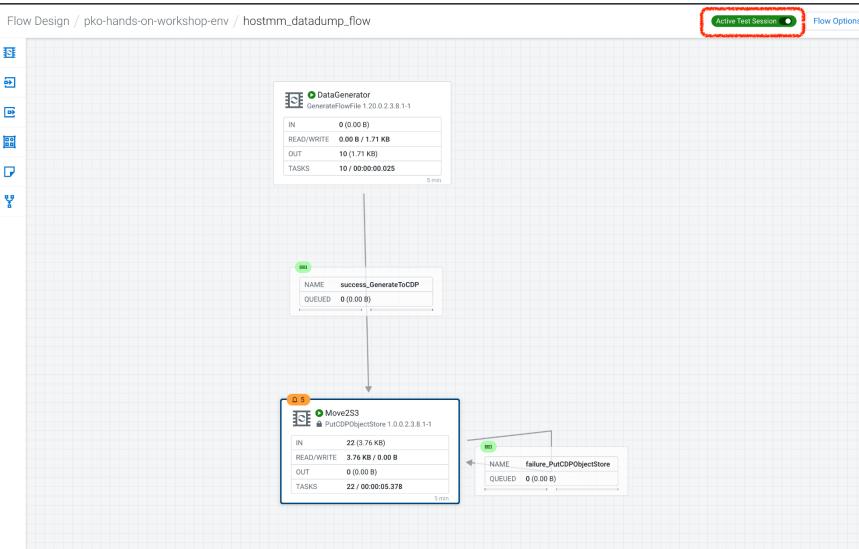
[You will not be able to access this S3 bucket by your self but the instructor will show you where everyones data is moving to]

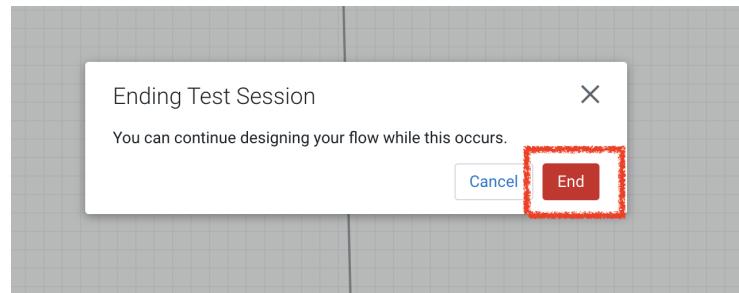
4. Move the Flow to the Flow Catalog

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

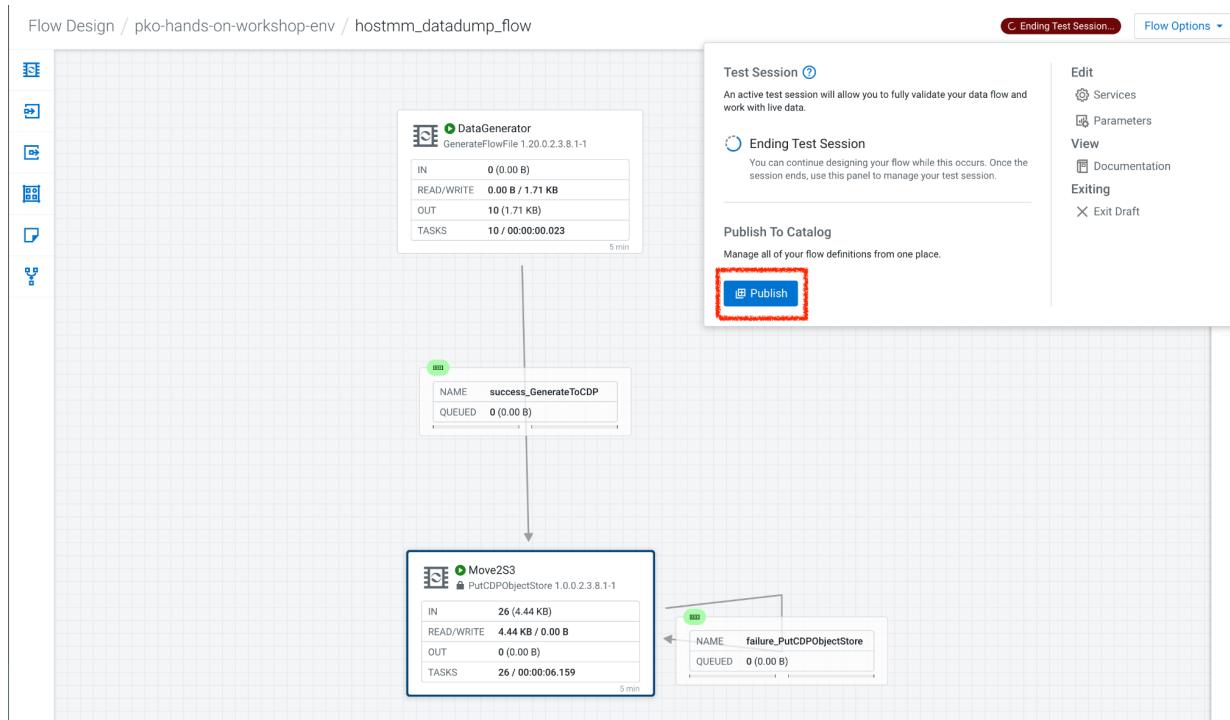
STOP the current test session by clicking on the green tab on top right and click END





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
DataDumpFlow

Flow Description 0/1K
Add description

Version Comments 15/1K
Initial Version

Cancel Publish

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

»

DataDumpFlow Updated 3 seconds ago by Manick Mehra

FLOW DESCRIPTION
No description specified
CRN # [crn:cdp:df:us-west-1:d1a4553c-a799-432d-8e54-372cc2ab95f2:flow:DataDumpFlow](#)

Only show deployed versions

Version	Deployments	Associated Drafts
1	0	1

Deploy → [Download](#) [Create New Draft](#)

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

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

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

Success!
Flow "DataDumpFlow" published successfully.

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 table header is labeled 'Name ↑'. Underneath, there is one row containing the text 'DataDumpFlow'. The entire interface is contained within a light gray box.

Click on the Flow, you should see the following:

The screenshot shows the details page for the 'DataDumpFlow' flow. At the top left is a small purple icon followed by the flow name 'DataDumpFlow'. To its right is a timestamp 'Updated 4 minutes ago by Manick Mehra'. On the far right is a 'Actions' dropdown menu. Below this, a section titled 'FLOW DESCRIPTION' contains the text 'No description specified' and 'CRN # 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 shows deployment statistics: Version 1 has 0 Deployments and 1 Associated Draft. At the bottom are buttons for 'Deploy →', 'Download', and 'Create New Draft'.

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 table with two columns: 'Version' and 'Deployments'. A red box highlights the first row, which contains the number '1' under 'Version' and '0' under 'Deployments'. Below this row is a blue button labeled 'Deploy →' with a red border around it. To the right of the button is a 'Download' link. Underneath the table, there are two sections: 'LAST UPDATE' (2021-09-23 11:52 CDT by Nasheb Ismaily, "Initial Version") and '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 screenshot shows a 'New Deployment' dialog box. At the top, it says 'Select the target environment'. Below this, there is a note: 'Sensitive data never leaves the environment. Changing the environment after this step requires restarting the deployment process.' The 'Selected Flow Definition' section shows a table with one row: NAME DataDumpFlow and VERSION 1. The 'Target Environment' section has a dropdown menu labeled 'Select an environment' with a blue border. Below the dropdown is a search bar labeled 'Filter by name'. A list of environments is shown: 'aws meta-workshop' (14% allocation), 'aws pko-hands-on-workshop-env' (15% allocation), and 'aws pse-workshop' (Workspace unavailable). The 'aws meta-workshop' row is highlighted.

Step 4: Deployment Name

Give the deployment a unique name (`{user_id}_flow_prod`), then click Next.
Example :
`wuser01_flow_prod`

Overview

Deployment Name
`flow tester`
 Deployment name is valid

Selected Flow Definition

NAME	VERSION
DataDumpFlow	1

Target Environment

aws	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: wuserXY

CDP Workload User Password:

Wuser@2021

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

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.

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

NiFi Node Sizing

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

Number of NiFi Nodes

Auto Scaling Enabled Disabled

Nodes:

Cancel **← Previous** **Next →**

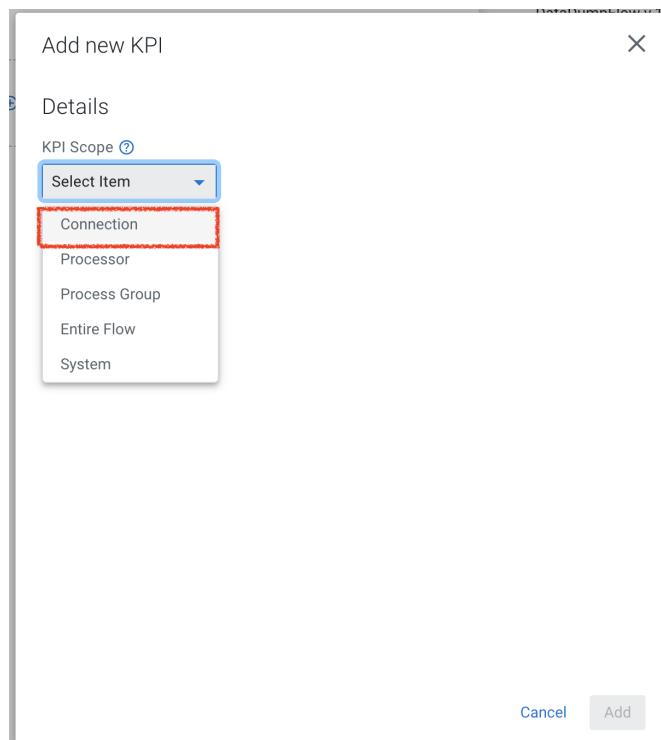
Step 8: Add Key Performance indicators

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

Click on “Add New KPI”

The screenshot shows a user interface for managing Key Performance Indicators (KPIs). At the top, the title "Key Performance Indicators" is displayed, followed by a subtitle: "Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed." Below this, there is a link "Learn more" with a blue arrow icon. In the center, there is a large, empty rectangular area with a dashed border, and in the middle of this area, there is a button labeled "+ Add New KPI". This central area is also outlined with a red box.

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



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

Add new KPI

Details

KPI Scope [?](#) Connection Name [?](#)

Connection failure_Move2S3

Metric to Track [?](#)

Percent Full

METRIC DESCRIPTION:
The percentage of connection that is full

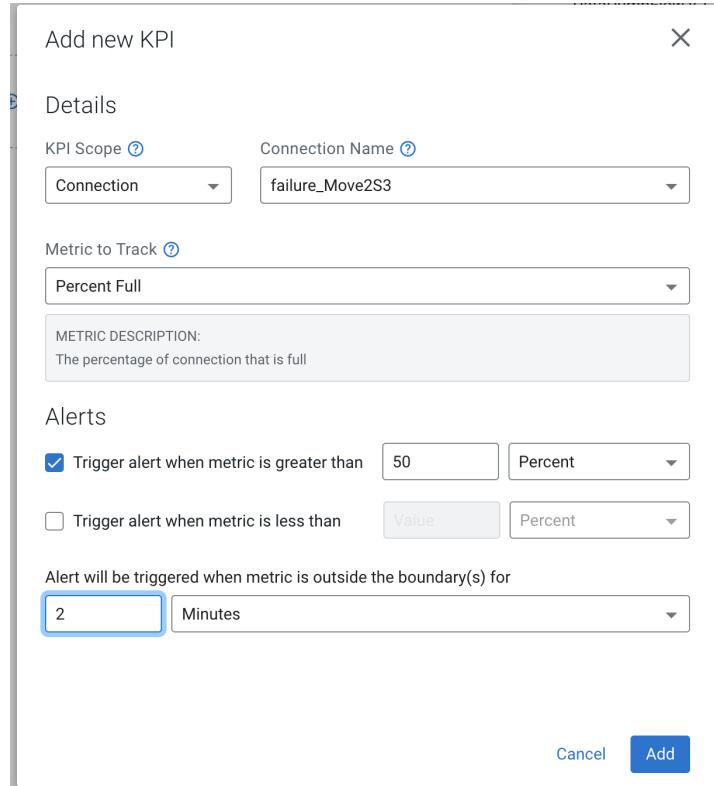
Alerts

Trigger alert when metric is greater than 50 Percent

Trigger alert when metric is less than Value Percent

Alert will be triggered when metric is outside the boundary(s) for 2 Minutes

Cancel Add



Click Add and then Click Next

Key Performance Indicators

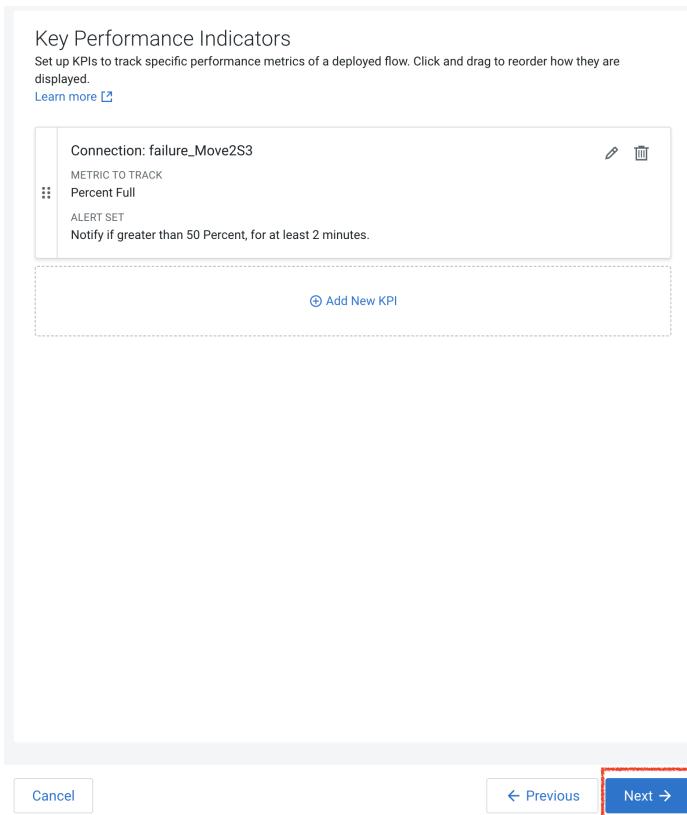
Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed.

[Learn more !\[\]\(2369187d55b5a244c8e464370db057ab_img.jpg\)](#)

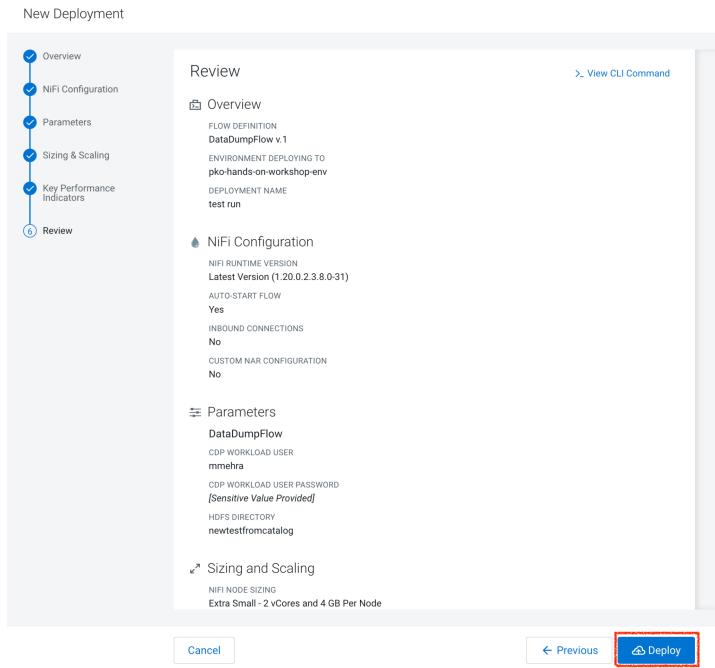
Connection: failure_Move2S3	 
METRIC TO TRACK	
Percent Full	
ALERT SET	
Notify if greater than 50 Percent, for at least 2 minutes.	

[+ Add New KPI](#)

Cancel [← Previous](#) **Next →**



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:

The screenshot shows the Data Flow Dashboard with the following filters applied: Filter By: STATUS All - 15, ENVIRONMENTS All - 4, DEPLOYMENTS All - 2, PROCESSOR TYPES All - 41, and METRICS WINDOW 30 Minutes. The dashboard lists three flows: gf-tbr-stocks-deployment (Suspended), hostmm_dataflow_prod (Good Health), and wuser00-syslog-to-kafka (Suspended). The hostmm_dataflow_prod flow is highlighted with a red box. Below the flows, there are two horizontal bar charts showing Data Throughput (Received/Sent) over 30 Minutes and Current. The hostmm_dataflow_prod chart shows 0 B/s for Received and 5 B/s for Sent. The bottom right corner of the dashboard interface has a red box around the 'Manage Deployment' button.

6. Viewing details of the deployed flow

Click on the flow in the Dashboard and select Manage Deployment

The screenshot shows the 'Deployment Information' section of the Data Flow Details page for the flow 'wuser122_flow_prod'. The flow was deployed by 'wuser122' on April 25, 2023, at 11:41 IST. It has an auto-scaling configuration of 'Up to 3 nodes'. The KPI tab is selected, showing a current percent full of 0.00% and an average of 0.00%. A boundary of > 50.00% is also indicated. The top right corner of the deployment information panel has a red box around the 'Manage Deployment' button.

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

REFRESHED: 27 seconds ago

[Actions ▾](#)

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
ENVIRONMENT aws pko-hands-on-workshop-env	REGION Asia Pacific (Mumbai)	NIFI RUNTIME VERSION 1.20.0.2.3.8.2-2	CRN # crm: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	Edit	Delete
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
ENVIRONMENT aws pko-hands-on-workshop-env	REGION Asia Pacific (Mumbai)	NIFI RUNTIME VERSION 1.20.0.2.3.8.2-2	CRN # crm:cdp:df:us-west-1:d1a4553c-a799-432d-8e54-372...

[Recreate Deployment CLI Command](#)

Deployment Settings

KPIs and Alerts Sizing and Scaling (highlighted with a red box) 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 Enabled

Min. Nodes	Max. Nodes
1	3

Step 3 : Manage Parameters

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

The screenshot shows the 'Parameters' tab of the Deployment Settings interface. It displays three parameter entries:

- wuser122_datadump_flow (4)**: CDP Workload User. Value: wuser122. Status: 8/100K.
- CDP Workload User Password**: Sensitive value provided. Status: 0/100K.
- CDPEnvironment**: DummyValue. Status: 10/100K.

Each entry includes checkboxes for 'Sensitive' and 'No value'.

Step 4 : NiFi Configurations

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

The screenshot shows the 'NiFi Configuration' tab of the Deployment Settings interface. It contains two sections:

- Inbound Connection Details**: A message states: "Inbound Connection has not been configured for this deployment."
- Custom NAR Configuration**: A message states: "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

Deployment Name		Flow Definition	Deployed By
wuser122_flow_prod	wuser122_datadump_flow V.1	wuser122	
Status	Node Count	Created On	Last Updated
Good Health	1	2023-04-25 11:41 IST	2023-04-25 11:43 IST
Environment	Auto Scaling	Nifi Runtime Version	CRN #
aws pk0-hands-on-workshop-env	Up to 3 nodes	1.20.0.2.3.8.2-2	cnr:cdp:df-us-west-2
Region			
Asia Pacific (Mumbai)			

Actions

- View in NiFi
- (i) Suspend flow
- (i) Change NIFI Runtime Version
- (i) 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

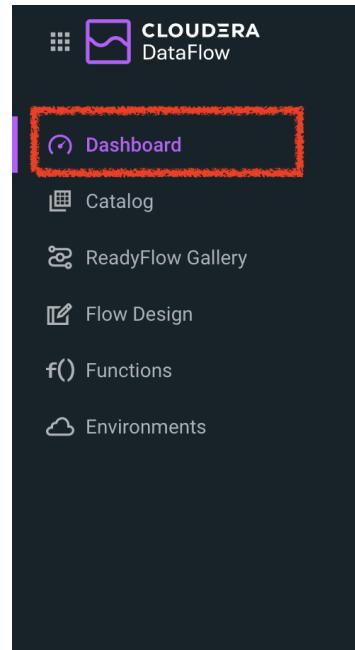
? Custom NAR has not been configured for this deployment.

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 refresh. Below the toolbar, a navigation bar includes links for 'Navigate' and search functions. The main workspace displays a flow diagram with a single node highlighted in blue. On the left, the 'Operate' panel shows a process named 'wuser122_flow_prod' with its ID as 'b70b3491-0187-1000-cf74-60ec3b41e85e'. It features a list of icons for managing the process. To the right, a detailed view of the flow 'wuser122_datadump_flow' is shown, listing metrics such as Queued (0 bytes), In (0 / 0 bytes), Read/Write (1.71 KB / 1.71 KB), and Out (0 → 0 bytes). Each metric has a corresponding status icon and a time period of 5 min.

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

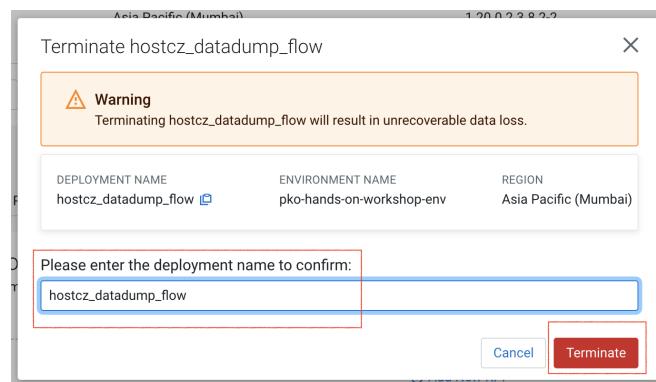
Select your flow and go to Manage Deployment

A screenshot of the Cloudera DataFlow UI dashboard. On the left, there's a sidebar with 'Dashboard', 'Catalog', 'ReadyFlow Gallery', 'Flow Design', 'Functions', and 'Environments'. The main area shows a table with a single row: 'hostcz_datadump_flow' (Status: Deploying). A red box highlights this row. To the right, there's a 'KPIs' section with a green circle and 'hostcz_datadump_flow pko-hands-on-workshop-env'. Below it is an 'Alerts' section with 'Active Alerts' (0) and 'No alerts to display.' A red arrow points from the 'Manage Deployment' link in the 'Alerts' section to the 'Manage Deployment' button at the bottom right of the page. The 'Event History' section shows deployment logs with a red box highlighting the first few entries. A 'Load More' button is at the bottom right of the event history table.

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, Deployed By, and Deployment Settings. On the right, there's an 'Actions' dropdown menu with options like View in NiFi, Suspend Flow, Change NiFi Runtime Version, Restart Deployment, and Terminate. A red box highlights the 'Actions' button and the 'Terminate' option.

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 tab. It displays a list of events under 'Event History'. The first event, 'Deployment Termination Initiated', is highlighted with a red box. Other events listed include 'Deployment Successful', 'NiFi Flow Started', 'KPI Alert Rules Activated', 'Activating KPI Alert Rules', 'Starting NiFi Flow', 'Default Alert Rules Activated', 'Activating Default Alert Rules', 'NiFi Flow Imported', and 'Importing NiFi Flow'. There are also sections for 'Active Alerts' (empty) and 'SHOW ONLY' filters for Info, Warning, and Error levels. 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:

- ssb-analytics-cluster
- kafka-smm-cluster

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 (kafka-smm-cluster)

The screenshot shows the Cloudera Management Console interface. On the left, there's a sidebar with various navigation options like Dashboard, Environments, Data Lakes, User Management, etc. The main area is titled 'Environments' and shows a cluster named 'pko-hands-on-workshop-env'. Below it, under 'Data Hub Clusters', there's a table listing four clusters: 'ssb-analytics-cluster', 'nifi-flow-mgmt-cluster', and 'kafka-smm-cluster' (which is highlighted with a red box). Each cluster entry includes its status (Running), type (e.g., Streaming Analytics Light Duty with Apache Flink), version (CDH 7.2.16), node count (6, 4, and 4 respectively), and creation date (04/20/23).

This screenshot shows the 'Services' page in the Cloudera Management Console. It lists several services: 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. The left sidebar has tabs for 'Topics' (which is highlighted with a red box) and 'Brokers'. The main area displays statistics for Producers (11) and Brokers (3). Below that, there's a table for Topics (30) and Brokers (3). The 'Topics' table shows a list of active topics with their names, sizes, and metrics. Some topics listed include '_consumer_offsets', '_CruiseControlMetrics', '_KafkaCruiseControlModelTrainingSamples', '_KafkaCruiseControlPartitionMetricSamples', '_smm_alert_notifications', and '_smm_consumer_metrics'.

3. Click on Add New

The screenshot shows the Apache Kafka UI's 'Topics' page. At the top, there are various metrics: Bytes In (1 MB), Bytes Out (917 KB), Produced Per Sec (2), Fetched Per Sec (1,784), In Sync Replicas (788), Out Of Sync (0), Under Replicated (0), and Offline Partitions (0). Below these are search and filter options. The main area is titled 'Topics (30)' and lists ten topics with columns for Name, Data In, Data Out, Messages In, Consumer Groups, and Current Log Size. The 'Add New' button in the top right corner 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

The screenshot shows the 'Add Topic' dialog box. It has fields for 'TOPIC NAME' (containing 'syslog') and 'PARTITIONS' (containing '1'). Below these are five availability options: Maximum, High, Moderate (selected), Low, and Custom. Under 'REPLICATION', it shows Factor 3, Min Insync Replicas 2 for Maximum; Factor 3, Min Insync Replica 1 for High; Factor 2, Min Insync Replica 1 for Moderate; Factor 1, Min Insync Replica 1 for Low; and Factor 1, Min Insync Replica 1 for Custom. A 'Limits' section shows a dropdown for 'CLEANUP.POLICY' set to 'delete'. At the bottom are 'Advanced', 'Cancel', and 'Save' buttons.

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.

3. Create a new schema with the following information:

- **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

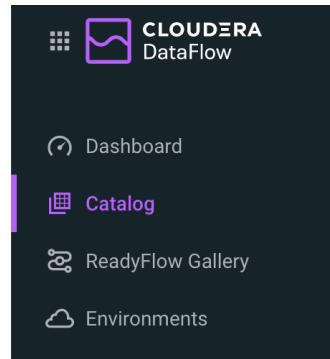
NAME *	host_cz-syslog	SCHEMA TEXT *
DESCRIPTION *	syslog schema for dataflow workshop	<input type="button" value="✖ CLEAR"/>
TYPE *	Avro schema provider	
SCHEMA GROUP *	Kafka	
COMPATIBILITY	BACKWARD	
<input checked="" type="checkbox"/> EVOLVE		<input type="button" value="CANCEL"/> <input style="border: 2px solid red;" type="button" value="SAVE"/>

SCHEMA REGISTRY		All Schemas			
		<input type="text" value="Search by name"/> <input type="button" value="🔍"/> Sort: Last Updated <input type="button" value="▼"/>			
NAME	TYPE	GROUP	BRANCH	SERIALIZER & DESERIALIZER	LAST UPDATED
hostmm_syslog	avro	Kafka	1 ↗	0	<input checked="" type="button" value="Schema added successfully"/>

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

Flow Name
syslog-to-kafka

Flow Description
Generates Syslog in RFC 5424 format, applies a SQL filter, transforms the data into JSON records, and publishes to Kafka
120/1000

NiFi Flow Configuration
syslog-to-kafka.json

Version Comments
Initial Version
15/1000

Click **IMPORT**

2. Deploy the Flow in CDF-PC

1. Search for the flow in the Flow Catalog

Flow Catalog

Q syslog-to-kafka

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

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**.

Version

Version	Deployments
1	0

Deploy →

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	VERSION
syslog-to-kafka	1

Target Environment

aws se-sandbox-aws	15% (3 of 20)	▼
--------------------	---------------	---

Cancel
Continue →

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

Example : {user_id}-syslog-to-kafka

Deployment Name

syslog-to-kafka

✓ Deployment name is valid

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

New Deployment

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

Cancel ← Previous Next →

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

- **CDP Workload User** - The workload username for the current user
 - Example : wuser00
 - **CDP Workload Password** - The workload password for the current user
[This password was set by you in Lab 0, section 3]
 - **Filtre Rule** - SELECT * FROM FLOWFILE
 - **Kafka Broker Endpoint** - A comma separated list of Kafka Brokers.
[Obtained in Lab 0, section 4]
- 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
```
- **Kafka Destination Topic** - <username>-syslog (Ex: wuser00-syslog)
  - **Kafka Producer ID** - nifi\_dfx\_p1
  - **Schema Name** - <username>-syslog (Ex: wuser00-syslog)
  - **Schema Registry Hostname** - The hostname of the master server in the Kafka Datahub(kafka-smm-cluster)[Refer screenshot below]

Data Hubs / kafka-smm-cluster / Nodes

| kafka-smm-cluster                                                                                                                                                                                                                 |                 |                                                            |                                                                                                                                              | Action            |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| STATUS                                                                                                                                                                                                                            | NODES           | CREATED AT                                                 | CLUSTER TEMPLATE                                                                                                                             | Stop Actions      |
| Running                                                                                                                                                                                                                           | 4 (0 0)         | 04/20/23, 11:33 AM GMT+5:30                                | 7.2.16 - Streams Messaging Light Duty: Apache Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager, Cruise Control |                   |
| STATUS REASON                                                                                                                                                                                                                     |                 |                                                            |                                                                                                                                              |                   |
| Cluster started.                                                                                                                                                                                                                  |                 |                                                            |                                                                                                                                              |                   |
| aws 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                                                                                                                                                                                                                             | Schema Registry | Streams Messaging Manager                                  | Token Integration                                                                                                                            |                   |
| Cloudera Manager Info                                                                                                                                                                                                             |                 |                                                            |                                                                                                                                              |                   |
| CM URL                                                                                                                                                                                                                            | CM VERSION      | RUNTIME VERSION                                            | LOGS                                                                                                                                         |                   |
| <a href="https://kafka-smm-cluster-gateway.pko-hand.dp5i-5vkq.cloudera.site/kafka-smm-cluster/cdp-proxy/conf/home/">https://kafka-smm-cluster-gateway.pko-hand.dp5i-5vkq.cloudera.site/kafka-smm-cluster/cdp-proxy/conf/home/</a> | 7.9.0           | 7.2.16-1.cdh7.2.16.p2.38683602                             | <a href="#">Command logs</a> , <a href="#">Service logs</a>                                                                                  |                   |
| Event History Autoscale Endpoints (5) Tags (4) Nodes (4) Network Load Balancers Telemetry Repository Details Image Details Recipes (0) Cloud Storage Database Upgrade Repair Delete                                               |                 |                                                            |                                                                                                                                              |                   |
| Master                                                                                                                                                                                                                            |                 |                                                            |                                                                                                                                              |                   |
| Instance ID                                                                                                                                                                                                                       | Status          | FQDN                                                       | Private IP                                                                                                                                   | Public IP         |
| i0db4b6eff7be50080                                                                                                                                                                                                                | Running         | kafka-smm-cluster-master0.pko-hand.dp5i-5vkq.cloudera.site | 10.10.220.8                                                                                                                                  | CM Server         |
| Broker                                                                                                                                                                                                                            |                 |                                                            |                                                                                                                                              |                   |

Example : kafka-smm-cluster-master0.pko-hand.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 [?](#)

|                                              |                                    |                                     |                                      |
|----------------------------------------------|------------------------------------|-------------------------------------|--------------------------------------|
|                                              |                                    |                                     |                                      |
| <input checked="" type="radio"/> Extra Small | <input type="radio"/> Small        | <input type="radio"/> Medium        | <input type="radio"/> Large          |
| 2 vCores Per Node<br>4 GB Per Node           | 4 vCores Per Node<br>8 GB Per Node | 8 vCores Per Node<br>16 GB Per Node | 16 vCores Per Node<br>32 GB Per Node |

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](#)

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

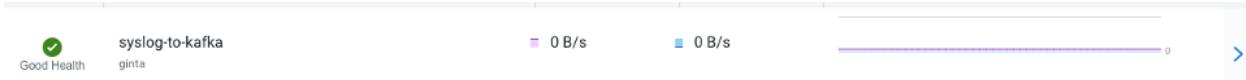
CUSTOM NAR CONFIGURATION  
No

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](#)

- 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.



- Click into your deployment and then Click **Manage Deployment** on the top right to view your flow in NiFi.

This screenshot provides a detailed view of the 'hostcz-syslog-to-kafka' deployment. On the left, the deployment table shows the flow definition 'hostcz-syslog-to-kafka V1' is suspended. On the right, the deployment information panel is expanded, showing details like node count (2), creation date (2023-04-27 08:21 IST), and last update (2023-04-27 20:11 IST). A red box highlights the 'Manage Deployment' button in the top right corner of this panel. Below the deployment information, there's a section for KPIs with a note: 'No KPIs to display. Set up key performance indicators to track specific aspects of your data flow to ensure it's operating as expected.' A 'Learn more' link is provided.

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

The screenshot shows the Deployment Manager page for a deployment named "hostcz-syslog-to-kafka V.1". Key details include:

- STATUS:** Good Health
- NODE COUNT:** 2
- ENVIRONMENT:** pko-hands-on-workshop-env
- FLOW DEFINITION:** hostcz-syslog-to-kafka V.1
- AUTO SCALING:** Up to 3 nodes
- CREATED ON:** 2023-04-27 08:21 IST
- REGION:** Asia Pacific (Mumbai)
- DEPLOYED BY:** host\_cz
- LAST UPDATED:** 2023-04-27 20:11 IST
- NIFI RUNTIME VERSION:** 1.20.0.2.3.8.2-2
- CRN #:** crn:cdp:df:us-west-2::hostcz-syslog-to-kafka

An "Actions" dropdown menu is open, showing options: View in NiFi, Suspend flow, Change NiFi Runtime Version, Restart Deployment, and Terminate. The "View in NiFi" option is highlighted with a red box.

**Deployment Settings**

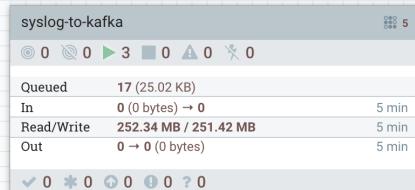
KPIs and Alerts   Sizing and Scaling   Parameters   NiFi Configuration

**Key Performance Indicators**

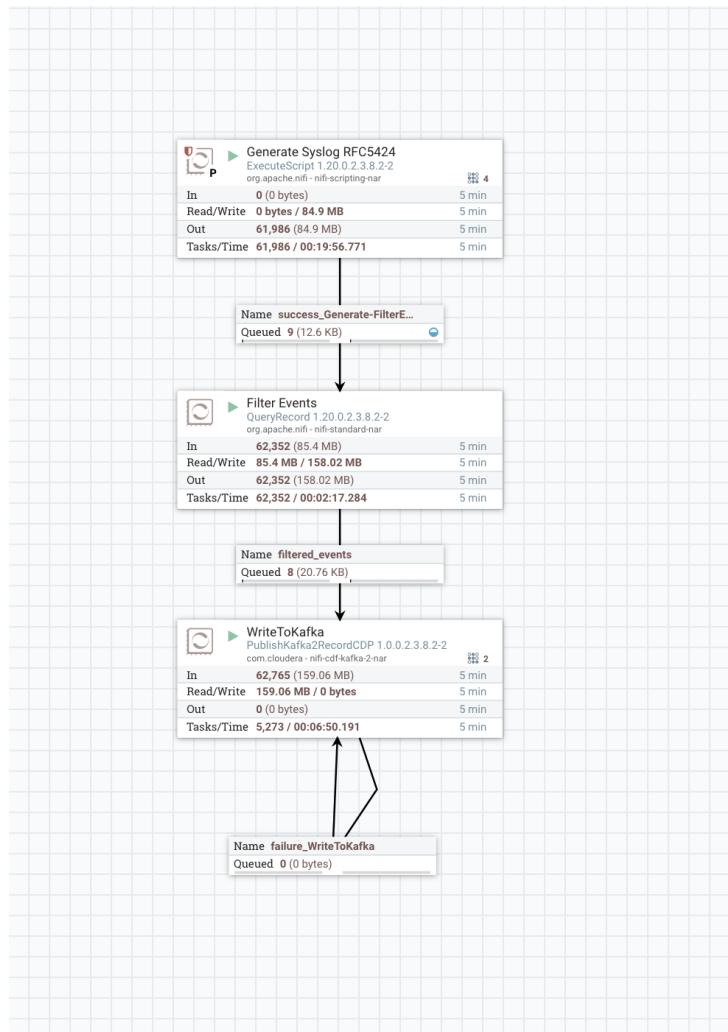
Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed.  
Learn more [\[?\]](#)

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

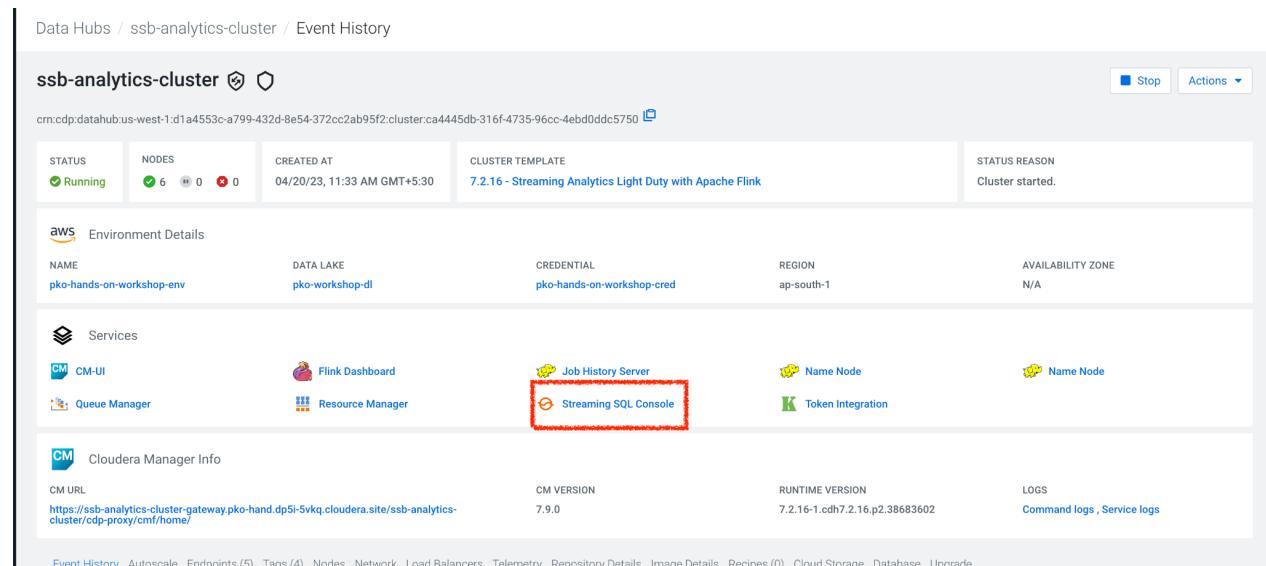
The purpose of this workshop is to demonstrate streaming analytic capabilities using Cloudera SQL Stream Builder. We will leverage the NiFi Flow deployed in CDF-PC from the previous lab and demonstrate how to query live data and subsequently sink it to another location. The SQL query will leverage the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

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

Below the summary card, there are sections for AWS Environment Details and Services. The Services section includes links for CM-UI, Flink Dashboard, Job History Server, Name Node, Token Integration, Queue Manager, Resource Manager, and Streaming SQL Console. The Streaming SQL Console link is highlighted with a red box. At the bottom, there's a Cloudera Manager Info section with a CM URL, CM Version (7.9.0), Runtime Version (7.2.16-1.cdh7.2.16.p2.38683602), and Logs (Command logs, Service logs). The footer contains links for Event History, Autoscale, Endpoints (5), Tasks (4), Nodes, Network, Load Balancers, Telemetry, Repository Details, Image Details, Recipes (0), Cloud Storage, Database, and Uninstall.

## 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 \*

hostmm\_hol\_workshop

Description

SSB Project to analysis streaming data

Override Materialized View Table Name Prefix ⓘ

Source Settings

Clone URL ⓘ

https://github.com/cloudera/ssb-examples.git

Branch ⓘ

main

Allow deletions on import ⓘ

Authentication

Create

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

hostmm\_hol\_workshop

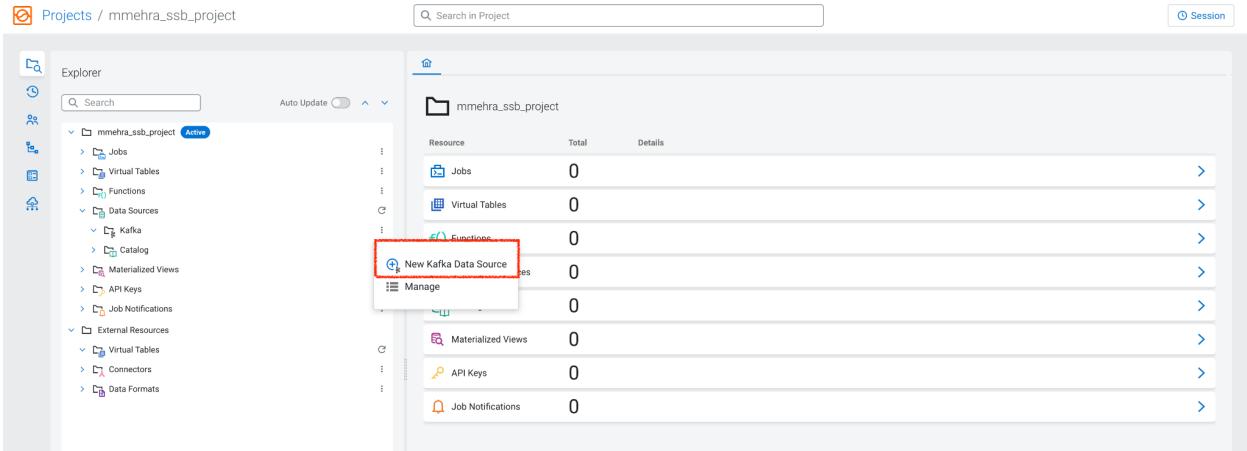
SSB Project to analysis streaming data

95d4615b

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)**

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

**Protocol :** SASL/SSL

**SASL Username :**

<workload-username>

Example : wuserXY

**SASL Password :** <Set in Lab

0 Section 3>

**SASL Mechanism :** PLAIN

**Kafka Data Source**

**Name \***  
njay-demo-ssb-kafka-ds

**Brokers (Comma-separated List) \***  
njay-demo-kafka-corebroker2.njay-dem.a465-9q4k.cloudera.site:9093,njay-demo-kafka-corebroker1.njay-dem.a465-9q4k.cloudera.site:9093,njay-demo-kafka-corebroker0.njay-dem.a465-9q4k.cloudera.site:9093

**Protocol \***  
SASL/SSL

**Kafka TrustStore (Optional)**  
/var/lib/cloudera-scm-agent/agent-cert/cm-auto-global\_truststore.jks

**Kafka TrustStore Password (Optional)**  
[redacted]

**Kafka KeyStore (Optional)**  
[redacted]

**Kafka KeyStore Password (Optional)**  
[redacted]

**SASL Mechanism**  
PLAIN

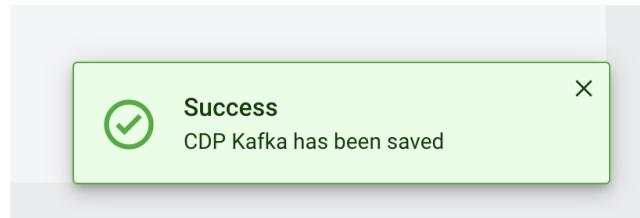
**Cancel** **Create**

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 Databricks UI interface. In the left sidebar, under the 'Virtual Tables' section, there is a three-dot menu icon. A red box highlights this menu icon. To its right, another red box highlights the 'New Kafka Table' option, which is shown as a blue button with a plus sign.

## 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 \*

Kafka Cluster \*

Data Format \*

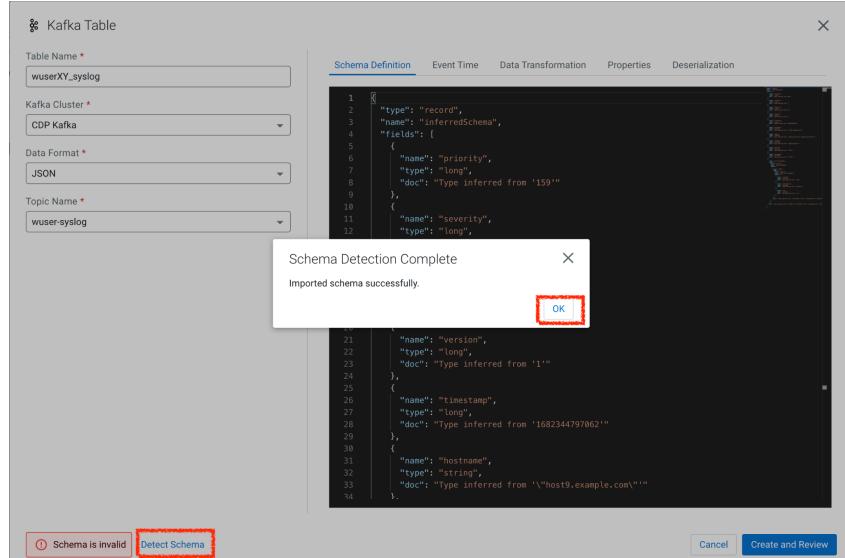
Topic Name \*

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.

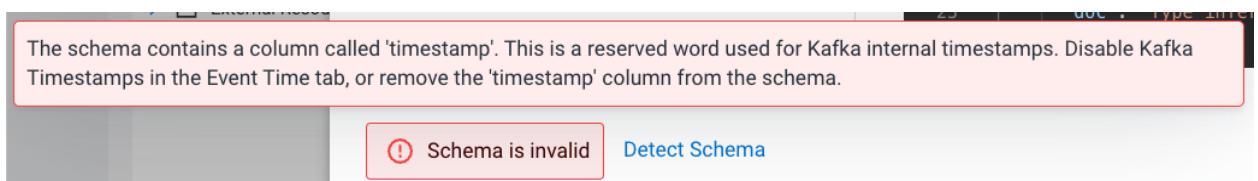
For JSON tables, though, SSB can look at the data flowing through the topic and try to infer the schema automatically, which is quite handy at times. Obviously, there must be data in the topic already for this feature to work correctly.

**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="float: left; margin-right: 10px;" type="button" value="Schema Definition"/> <input style="float: left; margin-right: 10px;" type="button" value="Event Time"/> <input style="float: left; margin-right: 10px;" type="button" value="Data Transformation"/> <input style="float: left;" type="button" value="Properties"/> <span style="float: right;">X</span> |                    |
| Input Timestamp Column: timestamp<br>Event Time Column: event_time<br>Watermark Seconds: 3<br>Use Kafka Timestamps: <input checked="" type="checkbox"/>                                                                                                                                                                                                                      |                    |

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

Schema is valid

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

Kafka Table

```

CREATE TABLE `syslog`.`host_test`.`wuser_syslog_kafka` (
 `priority` INT,
 `severity` INT,
 `facility` INT,
 `version` INT,
 `hostname` BIGINT,
 `hostname` VARCHAR(2147483647),
 `body` VARCHAR(2147483647),
 `appName` VARCHAR(2147483647),
 `appVersion` VARCHAR(2147483647),
 `messageId` VARCHAR(2147483647),
 `structuredData` ROW+ SOD+ ROW+ `eventId` VARCHAR(2147483647), `eventSource` VARCHAR(2147483647), `int` VARCHAR(2147483647)__,
 `eventTime` AS TO_TIMESTAMP_LTZ(`timestamp`, 3),
 `watermark` FOR `event_time` AS `event_time` - INTERVAL '3' SECOND
)
WITH (
 `properties.security.protocol` = 'SASL_SSL',
 `scm.startup.mode` = 'earliest-offset',
 `properties.bootstrap.servers` = 'localhost:9092',
 `properties.request.timeout.ms` = '120000',
 `properties.ssl.truststore.location` = '/var/lib/cloudera-scm-agent/agent/certs/cn-auto-global_truststore.jks',
 `properties.auto.offset.reset` = 'earliest',
 `properties.sasl.mechanism` = 'PLAIN',
 `format` = 'json',
 `properties.bootstrap.servers` = 'kafka-sem-cluster-corebroker1.pko-hand-5vqz.cloudera.site:9993,kafka-sem-cluster-corebroker2.pko-ha',
 `connector` = 'kafka',
 `properties.transaction.timeout.ms` = '900000',
 `topic` = 'syslog_test'
)

```

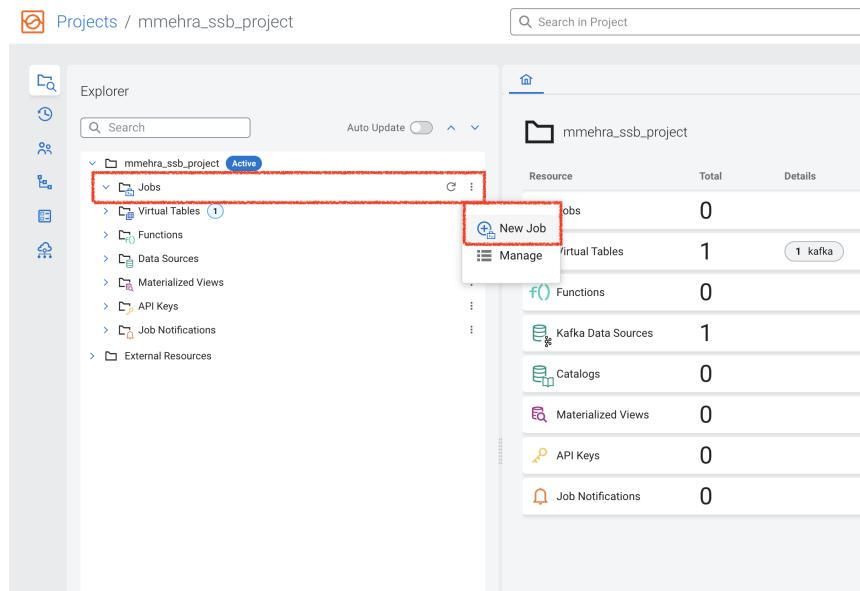
Success

wuser\_syslog\_kafka has been saved

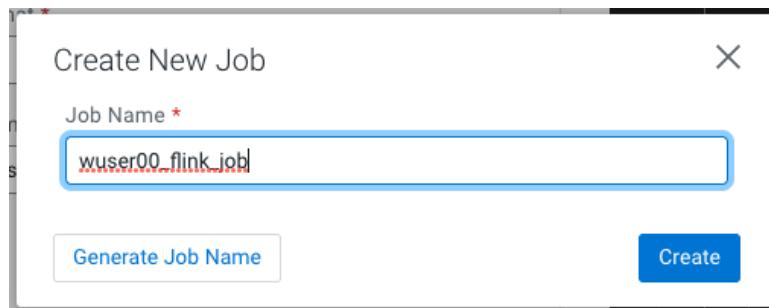
Review the table's DDL and click **Close**.

## Step 6: Create a Flink Job

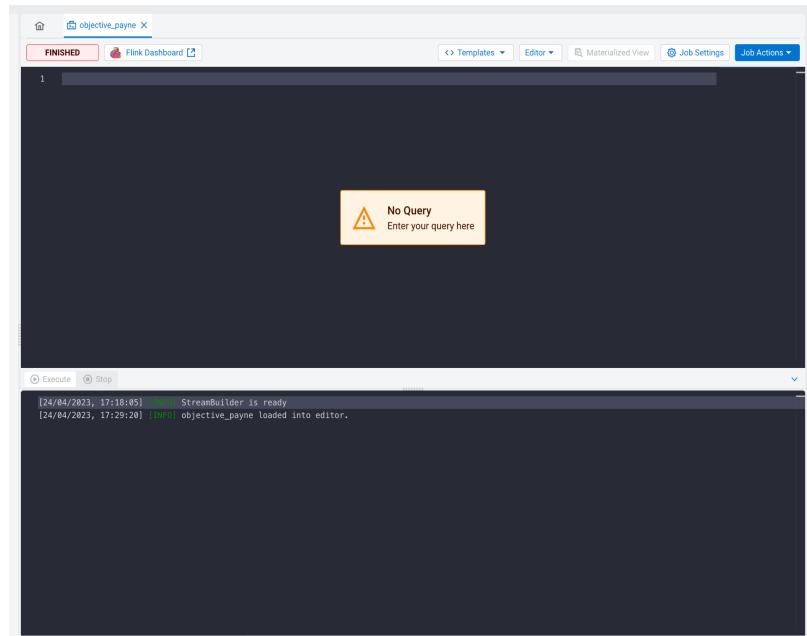
Create a Flink Job, by selecting “**Jobs**” in the left pane, clicking on the three-dotted icon next to it, then clicking on “**New Job**”.



Give a job name and click **CREATE**



The Query Editor should now show up

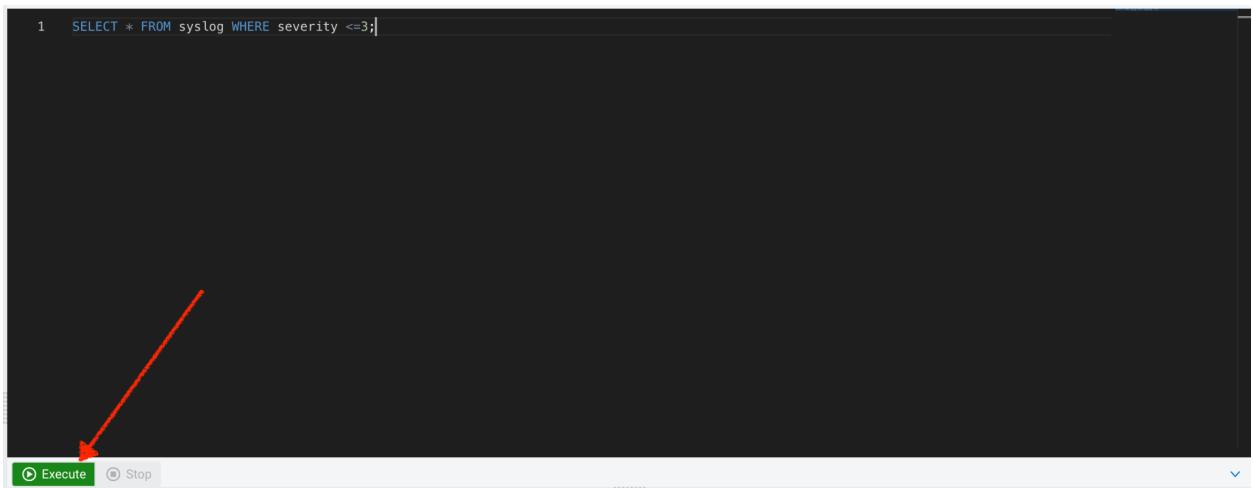


Add the following SQL Statement in the Editor

```
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

The screenshot shows the "Results" tab of a streaming SQL application. The results table has the following columns: pr..., severity, facility, version, timestamp, hostname, body, appName, procid, messageId, structured..., and event\_time. There are 10 rows of data displayed. Each row contains a checkbox, a severity level (e.g., 1, 2, 3, 4, 5, 6, 7, 10, 11, 13), a facility (e.g., 1b, 5, 4, 22, 13, 5, 21, 11), a version (e.g., 1, 1, 1, 1, 1, 1, 1, 1), a timestamp (e.g., 168243352..., 168243359..., 168243365..., 168243369..., 168243374..., 168243380..., 168243380..., 168243380...), a hostname (e.g., host1.exam..., host6.exam..., host7.exam..., host5.exam..., host6.exam..., host2.exam..., host4.exam..., host10.exa...), a body (application4..., application1..., application7..., application6..., application7..., application1..., application3...), an appName (application4..., application1..., application7..., application6..., application7..., application1..., application3...), a procid (2421, 1606, 1804, 7136, 1595, 6600, 7942, 3470), a messageId (IUS1, ID40, ID39, ID34, ID48, ID31, ID11, ID23), a structured... ({"SDID":("eve...), {"SDID":("eve...), {"SDID":("eve...), {"SDID":("eve...), {"SDID":("eve...), {"SDID":("eve...), {"SDID":("eve...), {"SDID":("eve...), {"SDID":("eve...), {"SDID":("eve...}), and an event\_time (2023-04-25..., 2023-04-25..., 2023-04-25..., 2023-04-25..., 2023-04-25..., 2023-04-25..., 2023-04-25..., 2023-04-25..., 2023-04-25..., 2023-04-25...). The "Logs" and "Events" tabs are also visible at the bottom of the interface.

| pr... | severity | facility | version | timestamp    | hostname      | body            | appName         | procid | messageId | structured...    | event_time    |
|-------|----------|----------|---------|--------------|---------------|-----------------|-----------------|--------|-----------|------------------|---------------|
|       | 1        | 1b       | 1       | 168243352... | host1.exam... | application4... | application4... | 2421   | IUS1      | {"SDID":("eve... | 2023-04-25... |
|       | 41       | 1        | 5       | 168243359... | host6.exam... | application1... | application1... | 1606   | ID40      | {"SDID":("eve... | 2023-04-25... |
|       | 34       | 2        | 4       | 168243365... | host7.exam... | application7... | application7... | 1804   | ID39      | {"SDID":("eve... | 2023-04-25... |
|       | 179      | 3        | 22      | 168243369... | host5.exam... | application6... | application6... | 7136   | ID34      | {"SDID":("eve... | 2023-04-25... |
|       | 106      | 2        | 13      | 168243374... | host6.exam... | application7... | application7... | 1595   | ID48      | {"SDID":("eve... | 2023-04-25... |
|       | 40       | 0        | 5       | 168243380... | host2.exam... | application1... | application10   | 6600   | ID31      | {"SDID":("eve... | 2023-04-25... |
|       | 170      | 2        | 21      | 168243380... | host4.exam... | application1... | application1... | 7942   | ID11      | {"SDID":("eve... | 2023-04-25... |
|       | 88       | 0        | 11      | 168243380... | host10.exa... | application3... | application3... | 3470   | ID23      | {"SDID":("eve... | 2023-04-25... |