

The Cloudera logo, featuring the word "CLOUDERA" in a bold, orange, sans-serif font.

CDE Workshop

Student Guide

Introduction	3
Prerequisites	4
Verify access to the environment	4
Download the resources	5
Update the files	7
Lab 1 - Walkthrough of CDE Data Service	7
ACCESSING THE VIRTUAL CLUSTER	11
Step 1 : Select the appropriate CDE Service	11
Step 2 : Virtual cluster selection	11
Lab 2 - Create and trigger ad-hoc Spark jobs	13
Resource Creation	13
Job Creation	15
Triggering the jobs	19
Lab 3 - Add schedule to the ad-hoc Spark jobs	21
Lab 4 - Orchestrate a set of jobs using Airflow	23
Lab 5 - Install and Configure CDE CLI	27
For Mac users:	27
For Windows users:	31
Lab 6 - Run jobs using CDE CLI	34
Run a spark-scala job using CLI	34
Lab 7 - Data Lineage and Auto-Scaling	35
Data Lineage using Atlas	35
Auto-scaling in CDE	38

Introduction

This document aims to introduce to our partners the features of **CDE**, the Data Engineering Data Service of Cloudera Data Platform (**CDP**). During the course of this workshop, you will experience how simple it is to run and orchestrate spark jobs with the help of auto-scaling infrastructure. We will use Airflow for orchestrating the various jobs.

In this workshop:

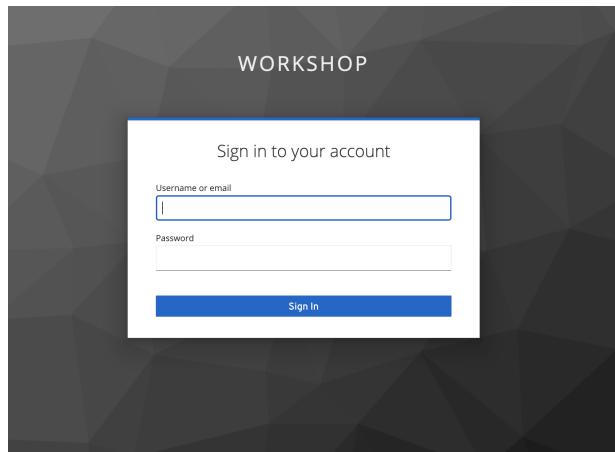
- You will be given an active CDE service running in an existing/registered CDP environment in a given tenant.
- You will have a virtual cluster with the given configuration that serves as compute for the spark workload.
- You will run sample spark jobs as ad-hoc jobs.
 - PySpark
 - Spark-scala
- You will run the same spark jobs as part of a schedule.
- With the help of Airflow, you will orchestrate a set of Spark jobs and trigger them as a flow.
- You will use CDE CLI to trigger the jobs from terminal/powershell.
- You will see the data lineage using Atlas and witness the auto-scaling capabilities of the CDE Data service.

Prerequisites

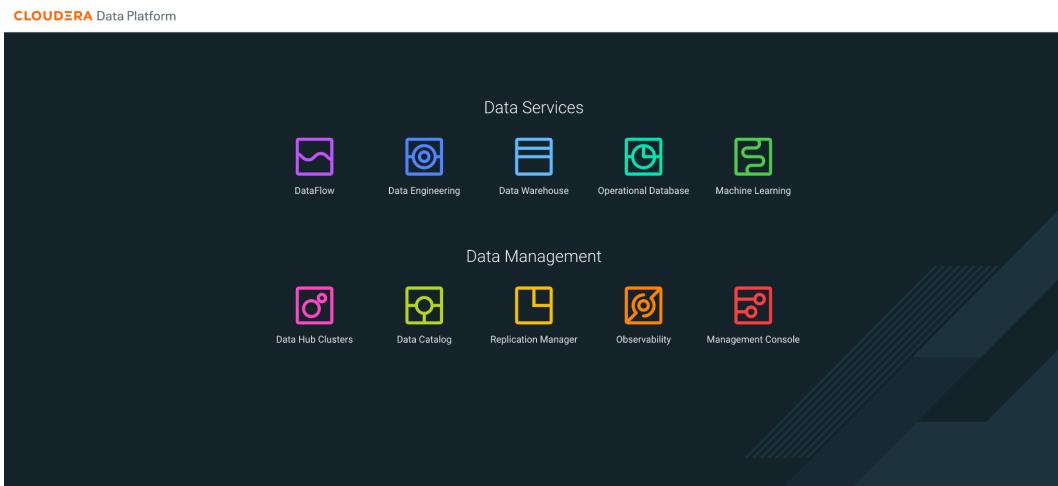
Verify access to the environment

- 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.

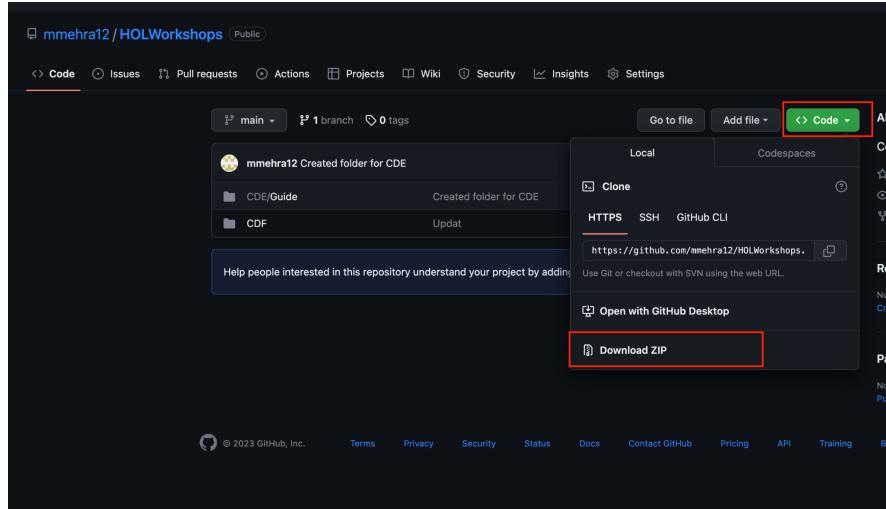


Download the resources

There are two ways in which you can access the scripts/resources.

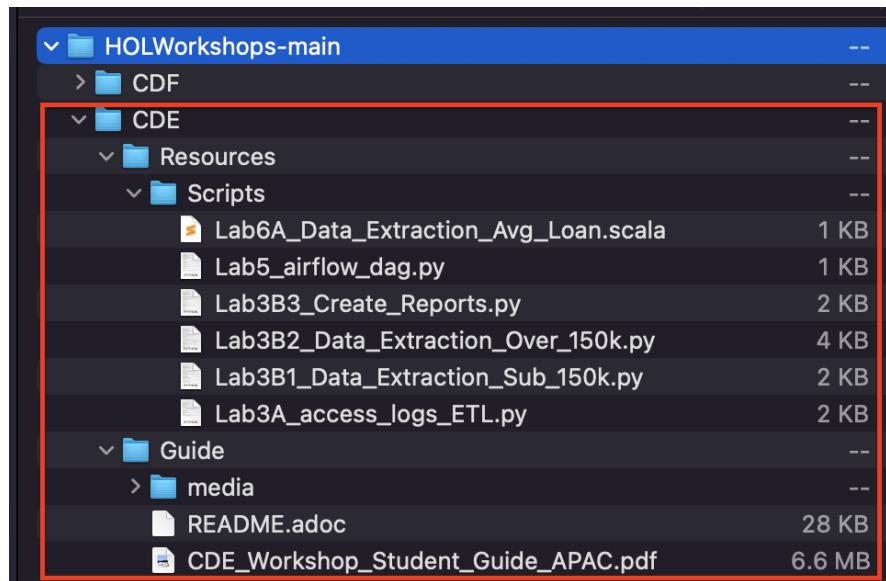
- Download the zip file from the GitHub repository.

<https://github.com/mmehra12/HOLWorkshops>



After decompressing the ZIP file the folder structure should look something like this

Note : We will use the CDE folder for this session, you can ignore the CDF content.

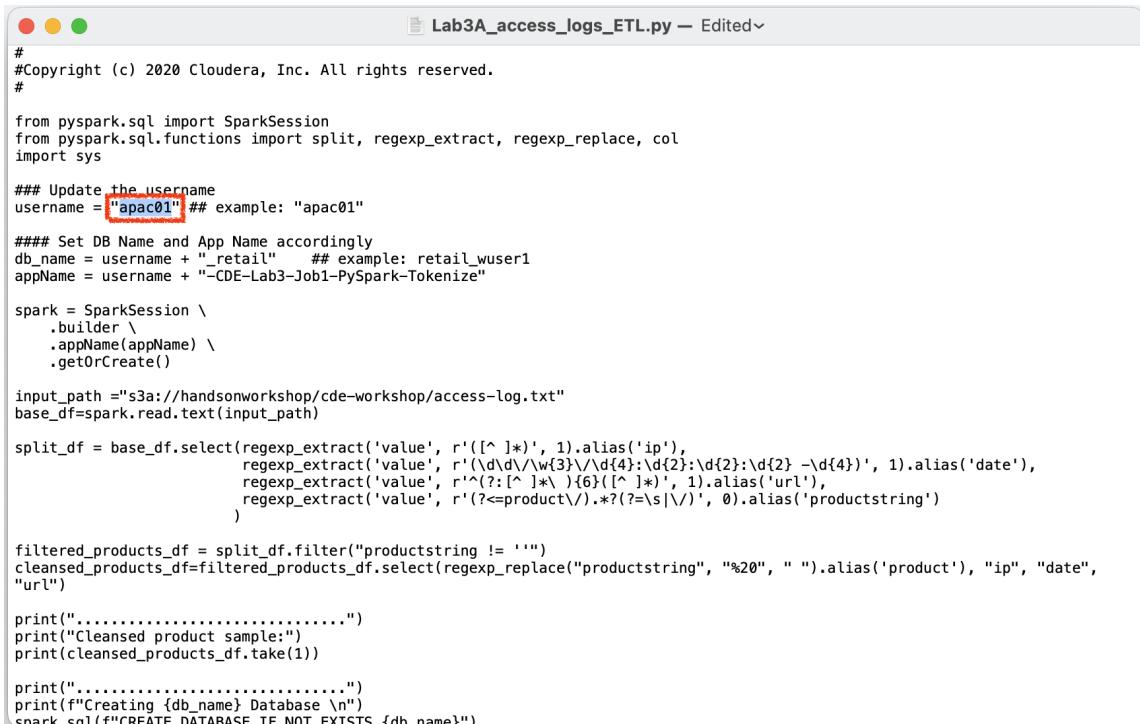


- The resources were also sent to you on your registered email an hour before the event. Please download the zip file attached to the email. After decompressing the ZIP file the folder structure should look something like this.

▼ CDE	--
▼ Resources	--
▼ Scripts	--
Lab5_airflow_dag.py	1 KB
Lab3A_access_logs_ETL.py	2 KB
Lab3B3_Create_Reports.py	2 KB
Lab3B2_Data_Extraction_Over_150k.py	4 KB
Lab3B1_Data_Extraction_Sub_150k.py	2 KB
Lab6A_Data_Extraction_Avg_Loan.scala	1 KB
▼ Guide	--
CDE_Workshop_Student_Guide_APAC.pdf	6.6 MB

Update the files

- Go through each script and update the necessary values as mentioned in the script.
 - For all the scripts, update the username field with the username that you have been assigned to. You will find this at the starting of the script itself.



```
#Copyright (c) 2020 Cloudera, Inc. All rights reserved.
#
from pyspark.sql import SparkSession
from pyspark.sql.functions import split, regexp_extract, regexp_replace, col
import sys
### Update the username
username = "apac01" ## example: "apac01"
#### Set DB Name and App Name accordingly
db_name = username + "_retail" ## example: retail_wuser1
appName = username + "-CDE-Lab3-Job1-PySpark-Tokenize"
spark = SparkSession \
    .builder \
    .appName(appName) \
    .getOrCreate()
input_path ="s3a://handsonworkshop/cde-workshop/access-log.txt"
base_df=spark.read.text(input_path)
split_df = base_df.select(regexp_extract('value', r'([^\s]+)', 1).alias('ip'),
                         regexp_extract('value', r'(\d\d\d\w{3}\d{4}\d{2}\d{2}\d{4})', 1).alias('date'),
                         regexp_extract('value', r'(?:[^\s]{6}[^\s]*)', 1).alias('url'),
                         regexp_extract('value', r'(?=<product\b).*(?=\s|\b)', 0).alias('productstring'))
filtered_products_df = split_df.filter("productstring != ''")
cleansed_products_df=filtered_products_df.select(regexp_replace("productstring", "%20", " ").alias('product'), "ip", "date",
"url")
print(".....")
print("Cleansed product sample:")
print(cleansed_products_df.take(1))
print(".....")
print(f"Creating {db_name} Database \n")
spark.sql(f"CREATE DATABASE IF NOT EXISTS {db_name}")
```

Lab 1 - Walkthrough of CDE Data Service

Cloudera Data Engineering (CDE) is a serverless service for Cloudera Data Platform that allows you to submit jobs to auto-scaling virtual clusters.

The CDE service involves several components:

- **Environment**
 - A logical subset of your cloud provider account including a specific virtual network.
- **CDE Data Service**
 - The long-running Kubernetes cluster and services that manage the virtual clusters. The CDE service must be enabled in an environment before you can create any virtual clusters.
- **Virtual Cluster**
 - An individual auto-scaling cluster with defined CPU and memory ranges. Virtual Clusters in CDE can be created and deleted on demand. Jobs are associated with clusters.
- **Job**
 - Application code along with defined configurations and resources. Jobs can be run on demand or scheduled.
- **Resource**
 - A defined collection of files such as a Python file or application JAR, dependencies, and any other reference files required for a job.
- **Job run**
 - An individual job run.

The above components can be accessed in the following ways:

- Go to the CDP console and click on Data Engineering.



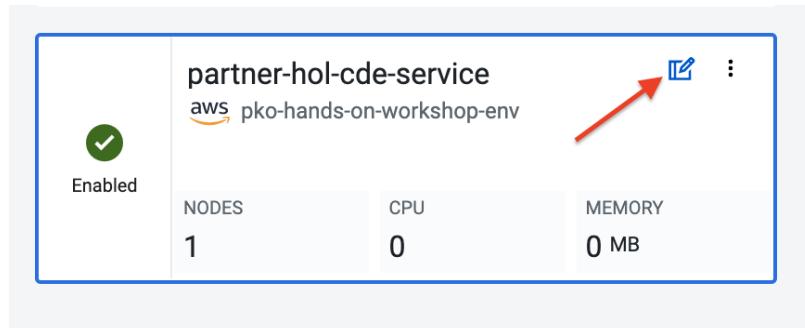
- You will see the CDE homepage.

Note : If the page load takes a while, you can move to the next step, we can come back to this later

- We should have a CDE service running which we will use for this workshop. To check this Select the **ADMINISTRATION** option on the left menu on your screen. You should be able to see all the CDE Service and their status.

Service	Status	Nodes	CPU	Memory
cdesrvc	Enabled	1	0	0 MB
cdetest	Failed	0	0	0 MB
meta-cde	Enabled	4	0	0 MB
partner-hol-cde-service	Enabled	1	0	0 MB

- On the **CDE service pko-workshop-cde-service**, click on the pencil icon and observe the configuration and other details related to the service.



Administration / Service / pko-workshop-cde-service

Enabled

pko-workshop-cde-service

VERSION	CLUSTER ID	CREATED BY	NODES	CPU	MEMORY
1.18.1-h3-b6	cluster-l87grkbn	Manick Mehra	1 / 50	0 / 400	0 MB / 1600 GB

GRAFANA CHARTS RESOURCE SCHEDULER

Configuration Logs Access Diagnostics

Environment: pko-hands-on-workshop-env

Workload Type: General - Small (EBS)

EBS Size: 100 GB

Capacity & Costs:
Autoscale Range: 1 to 50
On-demand Instances: 1 to 50

Use Spot Instances

Network & Storage:
 Enable Public Loadbalancer

API server Authorized IP Ranges: 0.0.0.0/0

Edit

- Click on each tab and go through all the details related to the CDE service.

- Once done, click on the **Home** on the left tab to go back to the CDE home page. This page shows us the active CDE services and the associate clusters. Let's start with accessing the virtual cluster that is assigned to you.

ACCESSING THE VIRTUAL CLUSTER

Step 1 : Select the appropriate CDE Service

Go to the Administration page and select your CDE Service (In our case **partner-hol-cde-service**)

The screenshot shows the Cloudera Data Engineering Administration interface. On the left, there is a sidebar with navigation links: Home, Jobs, Job Runs, Resources, and Administration. The Administration link is currently selected, indicated by a blue underline. The main content area is titled "Administration" and displays a list of CDE services:

CDE Service	Cluster	Status
cde-test	cdp-azure-demo-cdpwipdec	Failed
cdecdp	cdpoc	Failed
cdetest	cdpazure	Failed
meta-cde	aws meta-workshop	Enabled
pko-workshop-cde-ser...	aws pko-hands-on-workshop-env	Enabled
pse-workshop		

Below the table, resource statistics are shown for the "meta-cde" service: 4 nodes, 0 CPU, and 0 MB memory. The "pko-workshop-cde-ser..." service is highlighted with a blue border, indicating it is the selected service.

Step 2 : Virtual cluster selection

- Select the CDE Service and click on the virtual cluster that was assigned to you.

The screenshot shows the Cloudera Data Engineering (CDE) Administration interface. On the left, a sidebar lists navigation options: Home, Jobs, Job Runs, Resources, and Administration (which is selected and highlighted with a red border). Below the sidebar, it shows the user's name, Manick Mehra, and the IP address, 1.18.3-6.

The main area is titled "Administration" and displays a list of clusters:

- aws-cdtest-operation...**: Failed
- cde-test**: Failed
- cdecdp**: Failed
- cdetest**: Failed
- meta-cde**: Enabled (4 nodes, 0 CPU, 0 MB memory)
- pko-workshop-cde-ser...**: Enabled (highlighted with a blue border)

To the right of the cluster list, there is a summary card for the "hostcz-virtual-cluster" (pko-workshop-cde-service) showing 0 CPU, 0 MB memory, and 0 jobs. A red arrow points from the "Running" status of the pko-workshop-cde-service cluster towards this summary card.

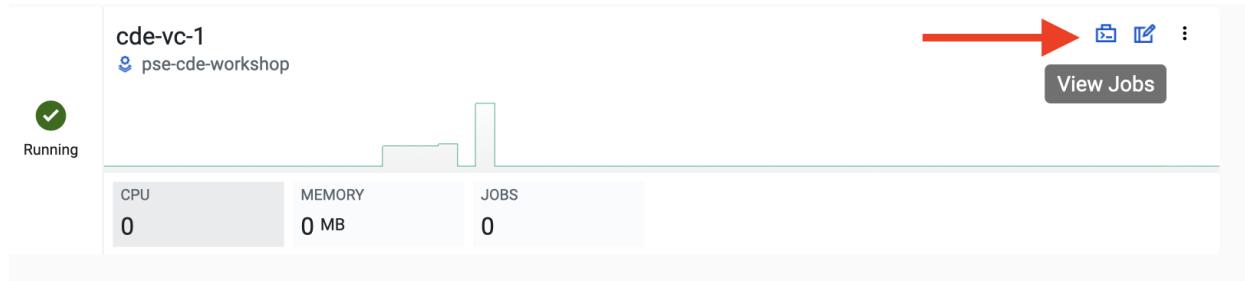
Lab 2 - Create and trigger ad-hoc Spark jobs

In this lab, we will create spark jobs and run them on an ad-hoc basis, i.e., without any schedule. As part of this lab, we have taken two simple use-cases that can be addressed with the help of Spark jobs.

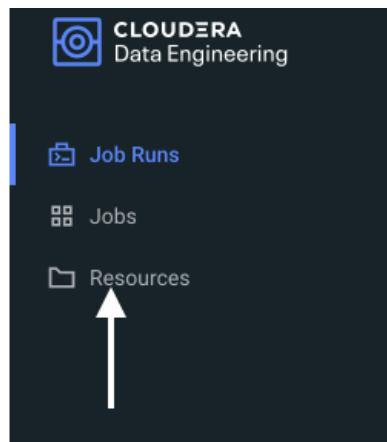
1. Log Data Cleansing using Spark
2. Analyze the Paycheck Protection Program Data
 - a. Report 1: Breakdown of all cities in Texas that retained jobs
 - b. Report 2: Breakdown of company type that retained jobs
3. PySpark job to enrich your data using an existing data warehouse

Resource Creation

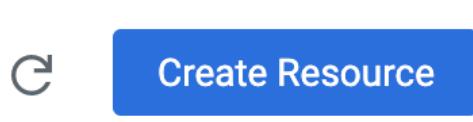
- On the virtual cluster **Cluster Name** : <username>-virtual-cluster [Virtual cluster created in Lab 1] tab, click on view jobs. This will open a new page with details of the Job Runs, Jobs, and Resources.



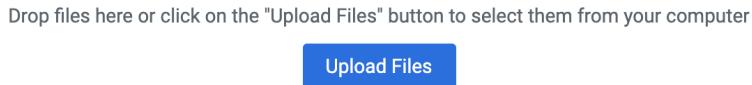
- In the left pane, click on the **Resources** tab.



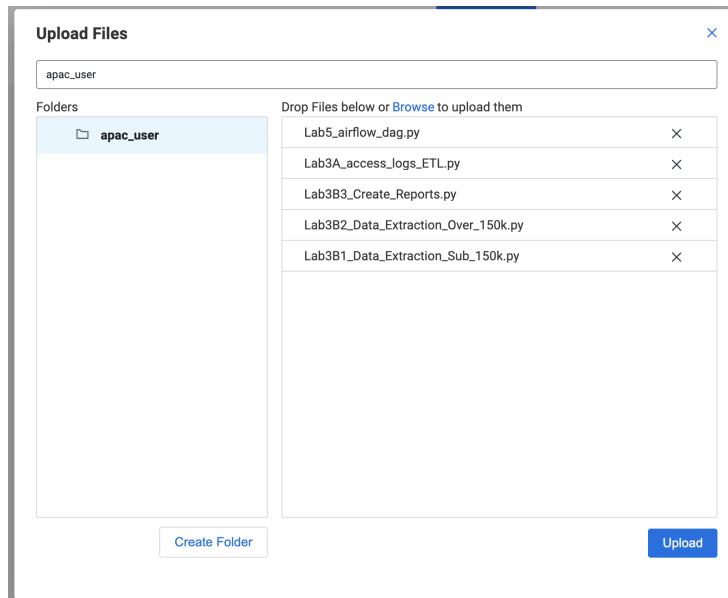
- You will get the **Resources** page to the right. Click on **Create Resource**.



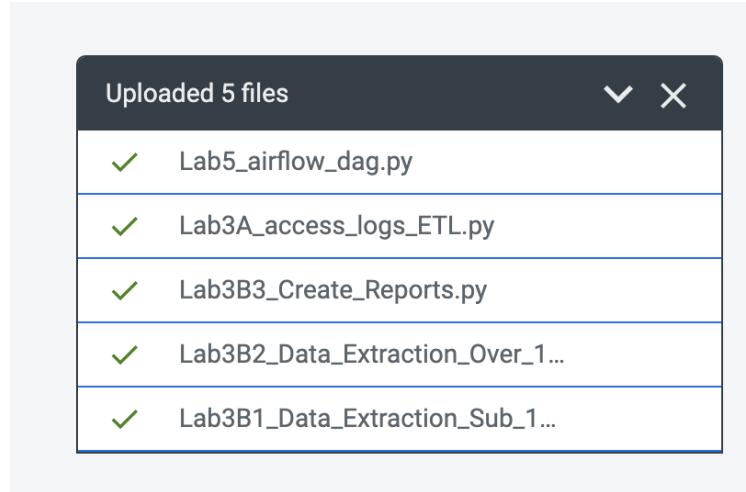
- Give a unique name(username-resources) and create the resource. This acts as your repository for storing all the scripts and dependencies.
- Once it is created, you will get an option to upload the files as shown below.



- Click on **Upload Files** and select all the scripts downloaded from the prerequisites step. (**Please upload only .py files**). Click on Upload



- You will get a pop-up with all the files uploaded to your resource.



- Validate if all the five .py files are present in your resource. We are now ready to create jobs using these resources.

Resources / apac_user / Files

vc-apac01

Name	Size	Modified On	Actions
Lab3A_access_logs_ETL.py	1.9 KIB	May 15, 2023, 12:01:42 PM	
Lab3B1_Data_Extraction_Sub_150k.py	2.2 KIB	May 15, 2023, 12:01:42 PM	
Lab3B2_Data_Extraction_Over_150k.py	3.7 KIB	May 15, 2023, 12:01:42 PM	
Lab3B3_Create_Reports.py	2.3 KIB	May 15, 2023, 12:01:42 PM	
Lab5_airflow_dag.py	1.4 KIB	May 15, 2023, 12:01:42 PM	

Upload Files

Items per page: 10 1 – 5 of 5 < >

Drop files or click on the "Upload Files" button to select them from your computer

Job Creation

- We will now create the first job with the script **Lab3A_access_logs_ETL.py**.
- In the left pane, click on **Jobs**.
- You will get the **Jobs** page to the right. Click on **Create Job**.

**Create Job**

- Select job type as **Spark**.
- Please give the job names as mentioned below.

<username>_<script_name_without_py_extension>

Eg:- For apac01, job1 name would be **apac01_Lab3A_access_logs_ETL**

Job Details

Job Type *

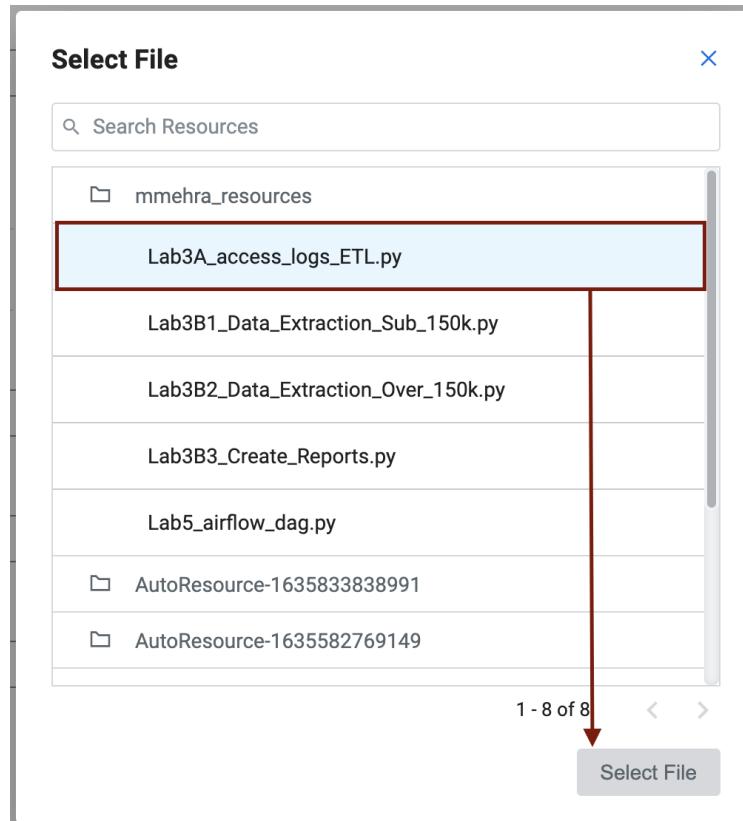
 Spark 2.4.7 Airflow

Name *

wuser01_Lab3A_access_logs_ETL

- As this is a shared environment, please name the jobs with your username so that it helps in differentiating yours from others' jobs.
- In Application File, click on **Select from Resource** and select the file **Lab3A_access_logs_ETL.py** from your resource(<username>-resources).

Application File File ⓘ URL ⓘUpload or **Select from Resource**



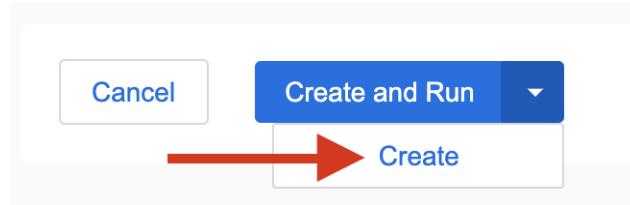
- Ignore the remaining configuration options. Do not enable the schedule now. This is how it should finally look like.

The screenshot shows the 'Jobs / Create Job' form in the CDE interface. On the left is a sidebar with 'Job Runs', 'Jobs' (which is selected), and 'Resources'. The main form has the following fields:

- Job Type ***: Spark 2.4.7 (radio button selected)
- Name ***: mmehra_Lab3A_access_logs_ETL
- Application File**:
 - File** (radio button selected)
 - URL (radio button unselected)A dropdown menu shows 'Lab3A_access_logs_ETL.py' with a checkmark.
- Arguments (Optional)**: An empty input field with a '+' icon.
- Configurations (Optional)**: Two empty input fields labeled 'config_key' and 'config_value' with a '+' icon.
- Python Version**: Python 3 (radio button selected)
- Python Environment**: A dropdown menu labeled 'Select Python Environment'.
- Advanced Options**: A collapsed section with the sub-instruction 'Upload additional files, customize no. of executors, driver and executor cores and memory'.
- Schedule**: A collapsed section with the sub-instruction 'Turn on to schedule Job, enable catchup and jobs dependants'.

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

- Click on the drop down option and click on **Create**. (do not click Create and Run)



- Similarly, create three other jobs with the same naming conventions. Please refer to the table below to confirm you are creating exactly the same.

For apac01:

Jobs	Job Name	Script Used
Job1	apac01_Lab3A_access_logs_ETL	Lab3A_access_logs_ETL.py
Job2	apac01_Lab3B1_Data_Extraction_Sub_150k	Lab3B1_Data_Extraction_Sub_150k.py
Job3	apac01_Lab3B2_Data_Extraction_Over_150k	Lab3B2_Data_Extraction_Over_150k.py
Job4	apac01_Lab3B3_Create_Report	Lab3B3_Create_Report.py

- Create these jobs as **ad-hoc** jobs i.e., without any schedule.
- Once done, click on the **Jobs** tab and enter your username in the search bar and press **ENTER**. You should see four jobs as shown below with your username.

Status	Job	Type	Schedule	Modified On	Actions
①	wuser01_Lab3B3_Create_Report	Spark	Ad-Hoc	Nov 3, 2021, 12:05:56 PM	⋮
①	wuser01_Lab3B2_Data_Extraction_Ov...	Spark	Ad-Hoc	Nov 3, 2021, 12:05:34 PM	⋮
①	wuser01_Lab3B1_Data_Extraction_Su...	Spark	Ad-Hoc	Nov 3, 2021, 12:05:09 PM	⋮
①	wuser01_Lab3A_access_logs_ETL	Spark	Ad-Hoc	Nov 3, 2021, 12:04:40 PM	⋮

- Observe the type of the job is set to Spark and for schedule, it is Ad-hoc.

Triggering the jobs

- You need to trigger the jobs in the following order
 - JOB 1 : apac01_Lab3A_access_logs_ETL
 - JOB 2 : apac01_Lab3B1_Data_Extraction_Sub_150k
 - JOB 3 : apac01_Lab3B2_Data_Extraction_Over_150k
 - JOB 4 : apac01_Lab3B3_Create_Reports(Run once JOB 2 and JOB 3 have completed successfully)

**NOTE : JOB 1, JOB 2 and JOB 3 can be triggered one after the other.
JOB 4 should be executed after the successful completion of JOB 2 and
JOB 3**

- To trigger the job, go to the **Jobs** tab, click on the 3-dotted icon, and click on **Run Now**.

Status	Job	Type	Schedule	Modified On	Actions
○	wuser01_Lab3B3_Create_Reports	Spark	Ad-Hoc	Nov 3, 2021, 12:05:56 PM	⋮
○	wuser01_Lab3B2_Data_Extraction_Ov...	Spark	Ad-Hoc	Nov 3, 2021, 12:05:34 PM	⋮
○	wuser01_Lab3B1_Data_Extraction_Su...	Spark	Ad-Hoc	Nov 3, 2021, 12:05:09 PM	⋮
○	wuser01_Lab3A_access_logs_ETL	Spark	Ad-Hoc	Nov 3, 2021, 12:04:40 PM	⋮

Items per page: 10 ▾

Run Now
Add Schedule
Clone
Configuration
Delete

- To check the job logs, click on **Job Runs** and select the ID against the job that you have triggered.

Status	Run ID	Job	Type	User	Duration	Start Time	Actions
○	47	wuser01_Lab3B2_Data_Extraction_Ov...	Spark	wuser01	Nov 3, 2021, 12:09:05 PM	2.4 MIN	⋮
○	46	wuser01_Lab3B1_Data_Extraction_Su...	Spark	wuser01	Nov 3, 2021, 12:09:04 PM	1.8 MIN	⋮
○	45	wuser01_Lab3A_access_logs_ETL	Spark	wuser01	Nov 3, 2021, 12:08:59 PM	1.9 MIN	⋮

Status	Run ID	Job	Type	User	Duration	Start Time	Actions
○	47	wuser01_Lab3B2_Data_Extraction_Ov...	Spark	wuser01	2.4 MIN	Nov 3, 2021, 12:09:05 PM	⋮
○	46	wuser01_Lab3B1_Data_Extraction_Su...	Spark	wuser01	1.8 MIN	Nov 3, 2021, 12:09:04 PM	⋮
○	45	wuser01_Lab3A_access_logs_ETL	Spark	wuser01	1.9 MIN	Nov 3, 2021, 12:08:59 PM	⋮

The screenshot shows the 'Job Runs / 47' page. At the top, it displays a summary for a job named 'wuser01_Lab3B2_Data_Extra...', which has a status of 'Succeeded', a duration of '2.4 min', and was started on 'Nov 3, 2021, 12:09:05 PM'. Below this, there are tabs for 'Trends', 'Configuration', 'Logs', and 'Spark UI'. A chart titled 'Duration' shows a single data point at 2.4 MIN. A search bar labeled 'Search by Run Id' is present. A table below lists the job details: Status (Succeeded), Run ID (47), Duration (2.4 min), Executor Memory (1G), Drivers Cores (1), Executor Cores (1), User (wuser01), and Start Time (Nov 3, 2021, 12:09:05 PM). An 'Actions' column contains a three-dot menu icon.

- For simplifying the job selection, you can choose the **User** filter and add your username and hit enter. You will see the list of jobs triggered by you.

Job Runs

The screenshot shows the 'Search Job Runs' interface. It includes a search bar with a magnifying glass icon, a dropdown menu set to 'User', and a 'Search Job Runs' button. Below the search bar, there are filters for 'Run Id', 'Status', 'Job Name', and 'User'. The 'User' filter is highlighted with a gray background and blue text. A large red arrow points to the 'User' dropdown menu.

- Navigate to different tabs in the job run page and you will see all that you need to observe for the run of a Spark job.

The screenshot shows the navigation bar for the job run page. It features four tabs: 'Trends' (which is selected and highlighted in blue), 'Configuration', 'Logs', and 'Spark UI'. Each tab has a corresponding icon above it: a line graph for Trends, a gear for Configuration, a document for Logs, and a spark icon for Spark UI.

Lab 3 - Add schedule to the ad-hoc Spark jobs

In this lab, we will add a schedule to a job created as part of the previous lab.

- We will add a schedule to the job **Lab3A_access_logs_ETL** (in your case it will be <username>_Lab3A_access_logs_ETL)
- Go to **Jobs** tab, click on the 3-dotted icon next to the job **Lab3A_access_logs_ETL** and select **Add schedule**.

The screenshot shows the Cloudera Data Engineering interface with the 'Jobs' tab selected. A list of jobs is displayed, including 'hostcz_Lab3B3.Create_Reports', 'hostcz_Lab3B2.Data_Extraction_Over_150k', 'hostcz_Lab3B1.Data_Extraction_Sub_150k', and 'hostcz_Lab3A.access_logs.ETL'. A context menu is open over the last job, with the 'Add Schedule' option highlighted by a red box and arrow. Other options in the menu include 'Run Now', 'Clone', 'Configuration', and 'Delete'.

- You will land in the **Job Schedule** page. Click on **Create a Schedule**.

The screenshot shows the 'Job Schedule' page for the 'hostcz_Lab3A.access_logs.ETL' job. The 'Schedule' tab is selected. A message at the top states 'This job currently does not have a schedule.' Below this, there is a 'Create a Schedule' button. The page also includes tabs for 'Run History' and 'Configuration'.

- Choose the **Cron Expression** option and enter the cron expression as given below.

***/10 * * * *** → This means that the job is scheduled to run every 10 minutes.

The screenshot shows the CDE interface for scheduling a job. The 'Schedule' tab is selected. The cron expression is set to `*/10 * * * *`. The start date is Thursday, May 11, 2023 at 11:35:28 AM, and the end date is Friday, May 12, 2023 at 11:35:28 AM. There are two configuration options under 'Scheduling Configurations': 'Enable Catchup' (unchecked) and 'Depends on Previous' (unchecked). At the bottom right are 'Cancel' and 'Add Schedule' buttons. A green success message box is centered below the schedule form.

- You can repeat the same process for the other jobs as well.
 - JOB 1 : Run every 10 mins
 - JOB 2 : Run every 10 mins
 - JOB 3 : Run every 10 mins
 - JOB 4 : Run every 30 mins
- We do not have to wait for the jobs to get triggered as per the schedule. The idea was to understand how Ad-Hoc jobs are scheduled. We can continue with the next steps
- Please **PAUSE** the schedule for all the jobs for which it was added by following the below steps.
- Go to the Jobs tab, click on the 3-dotted icon next to the job and select **Pause schedule**. [Do this for all jobs]

The screenshot shows the CDE Jobs tab. It lists several jobs:

Status	Job	Type	Schedule	Modified On	Actions
○	skillupuser_Lab3A_access_logs_ETL	Spark	*/10 * * * *	May 11, 2023, 11:36:58 AM	⋮
○	hostcz_Lab3B1_Create_Report	Spark	*/10 * * * *	May 10, 2023, 11:36:58 AM	⋮
○	hostcz_Lab3B2_Data_Extraction_Over_150k	Spark	*/10 * * * *	May 10, 2023, 11:36:58 AM	⋮
○	hostcz_Lab5_airflow_dag	Airflow	*/10 * * * *	May 9, 2023, 12:31:33 PM	⋮
○	hostcz_Lab3A_access_logs_ETL	Spark	Ad-Hoc	May 9, 2023, 12:31:33 PM	⋮
○	hostcz_Lab3B1_Data_Extraction_Sub_150k	Spark	Ad-Hoc	May 9, 2023, 12:31:33 PM	⋮

 The 'Actions' column for the first job shows a context menu with options: Run Now, Pause Schedule (highlighted with a red box), Edit Schedule, Remove Schedule, Clone, Configuration, and Delete.

A green success message box is centered below the table:

The schedule for skillupuser_Lab3A_access_logs_ETL job has been paused.

Lab 4 - Orchestrate a set of jobs using Airflow

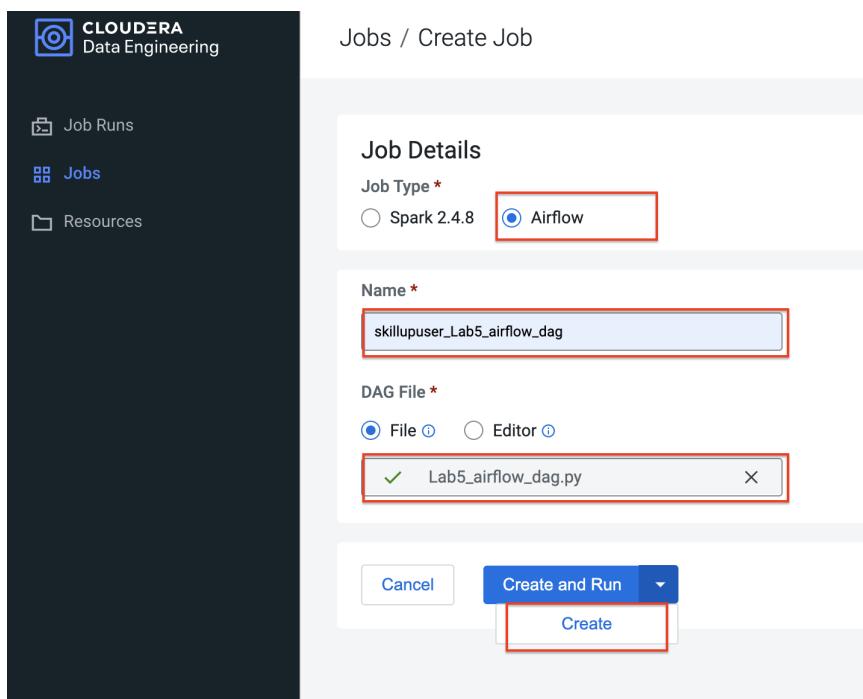
In this lab, we will create a flow with the help of a dag file that uses the jobs created in Lab3. Thus, you will be able to complete subsequent labs only if you have completed Lab3 successfully.

- Go to Jobs tab, click on **Create Job** and choose Airflow in Job type.
- Give the job name as below and upload the *Lab5_airflow_dag.py* file from the resources.

JOB NAME : <username>_Lab5_airflow_dag

Example : For user **apac01** the job name will be, **apac01_Lab5_airflow_dag**

- Click on **Create**.



- Go to **Jobs** tab and observe the airflow job created with the schedule mentioned in the dag file.

Job

Jobs		
Status	Job	Type
	hostcz_Lab5_airflow_dag	Airflow

DAG File

```

## Update the username
owner = "<ENTER YOUR USER NAME HERE>" # Example: "apac01"

DAG_name = owner + "_Airflow_Dag"
job_name_1 = owner + "_Lab3B1_Data_Extraction_Sub_150k"
job_name_2 = owner + "_Lab3B2_Data_Extraction_Over_150k"
job_name_3 = owner + "_Lab3B3_Create_Reports"

default_args = {
    'owner': owner,
    'retry_delay': timedelta(seconds=5),
    'email_on_failure': False,
    'email_on_retry': False,
    'retries': 0
}

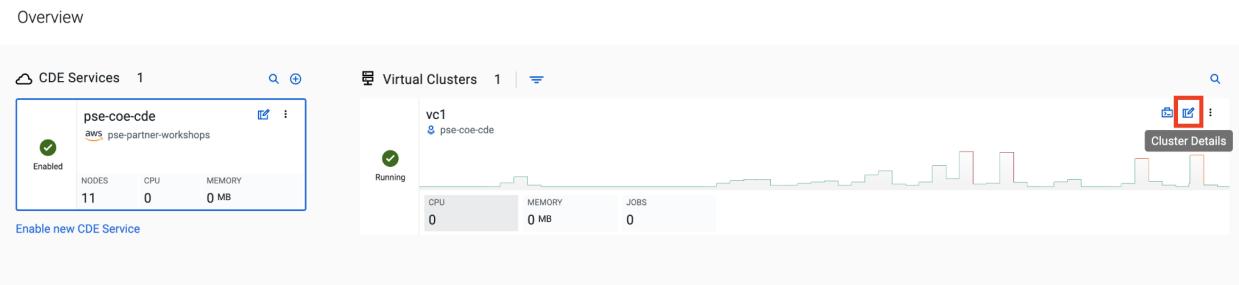
dag = DAG(
    DAG_name,
    default_args=default_args,
    start_date=datetime(2023,5,3),
    end_date=datetime(2023,5,18),
    schedule_interval='*/20 * * * *',
    catchup=False,
    is_paused_upon_creation=False
)

start = DummyOperator(task_id='start', dag=dag)

Data_Extraction_Sub_150k = CDEJobRunOperator(
    task_id=job_name_1,
    retries=3,
    dag=dag,
    job_name=job_name_1
)

```

- Go to the Virtual Cluster you are using and click on **Cluster Details**.



- Click on **Airflow UI** and observe the schedule created for your job.

Overview / vc1

vc1

VERSION 1.12.0-b119 VC ID dex-app-hg98mkj CREATED BY Panag Katti CPU 3 MEMORY 4 GB JOBS 0

ENVIRONMENT DATA LAKE

AIRFLOW UI

Configuration Charts Logs

CDE Service pse-partner-workshops

Autoscale Max Capacity

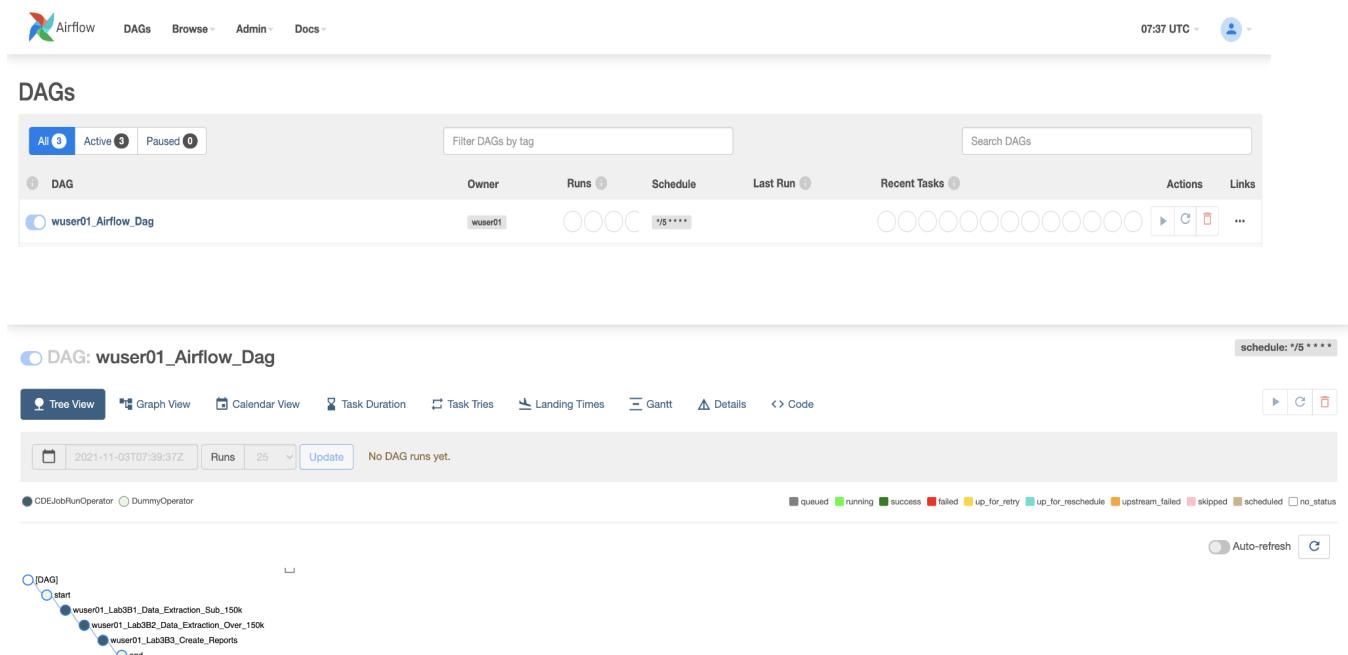
CPU: 30 to 600 (30)

Memory (GB): 4 to 2400 (508)

Driver and Executors will run on

Spark Version: Spark 2.4.7

Enable Airflow Job Authoring UI (Technical Preview)



- Once the job has run successfully, we need to edit the job to **pause** the schedule.
- Click on the Jobs tab and locate the airflow job that you have just created.
- Next to the job, click on the 3 dots and click on **Pause Schedule**.

The screenshot shows the CDE Jobs interface. At the top, there's a search bar, filter dropdowns for Status and Type, and a 'Create Job' button. Below is a table with columns: Status, Job, Type, Schedule, Modified On, and Actions. The 'Actions' column contains three dots, Run Now, Pause Schedule, Configuration, and Delete. The 'Pause Schedule' option is highlighted with a red box. A green success message box is overlaid on the interface, stating: 'The schedule for hostcz_Lab5_airflow_dag job has been paused.'

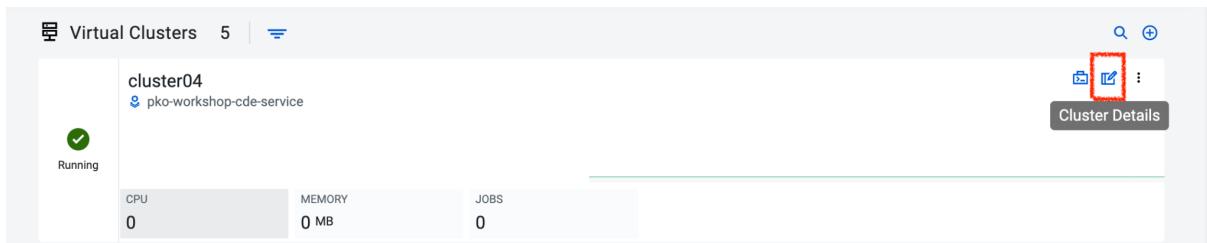
Status	Job	Type	Schedule	Modified On	Actions
○	hostcz_Lab5_airflow_dag	Airflow	*/20 * * * *	May 11, 2023, 11:43:07 AM	⋮ Run Now Pause Schedule Configuration Delete
○	skillupuser_Lab3A_access_logs_ETL	Spark	*/10 * * * *	May 11, 2023	⋮
○	hostcz_Lab3B3_Create_Reports	Spark	*/10 * * * *	May 10, 2023	⋮ Configuration Delete
○	hostcz_Lab3B2_Data_Extraction_Over_150k	Spark	*/10 * * * *	May 10, 2023, 4:01:34 AM	⋮

- You can go to the AirFlow UI again and see that the Job is now in Paused State

The screenshot shows the AirFlow UI with a single DAG listed: 'apac52_Airflow_Dag'. Above the DAG name, a status message 'DAG is Paused' is displayed. The DAG icon is greyed out, indicating it is currently paused.

Lab 5 - Install and Configure CDE CLI

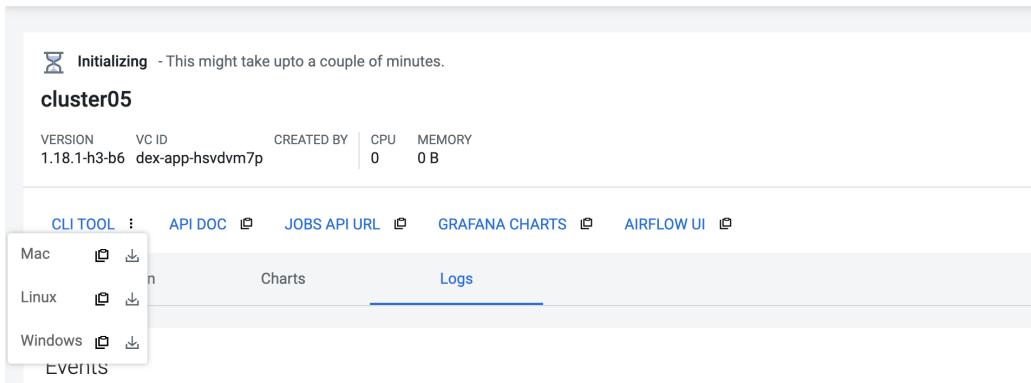
- In this lab, we will use the CDE CLI to create and run a spark job. This way, you can use the rich api's of CDE CLI to integrate any of your applications to communicate with the CDE service.
- The CLI executable can be downloaded from the virtual cluster.
 - **Step 1 :** Go to the **Cluster Details** of the virtual cluster where you are creating your job



The screenshot shows the Cloudera Data Engine (CDE) interface. On the left, there's a sidebar with a search icon and a plus sign. The main area is titled "Virtual Clusters" and shows 5 clusters. One cluster, "cluster04", is highlighted and has a green checkmark icon next to it, indicating it is running. Below the cluster name, it says "pko-workshop-cde-service". To the right of the cluster details, there are three icons: a magnifying glass, a gear, and a three-dot menu. Below these icons is a "Cluster Details" button, which is also highlighted with a red box. At the bottom of the cluster card, there are metrics: CPU 0, MEMORY 0 MB, and JOBS 0.

- **Step 2 :** Click on CLI TOOL to download the executable based on your operating system.

Administration / Virtual Cluster / cluster05



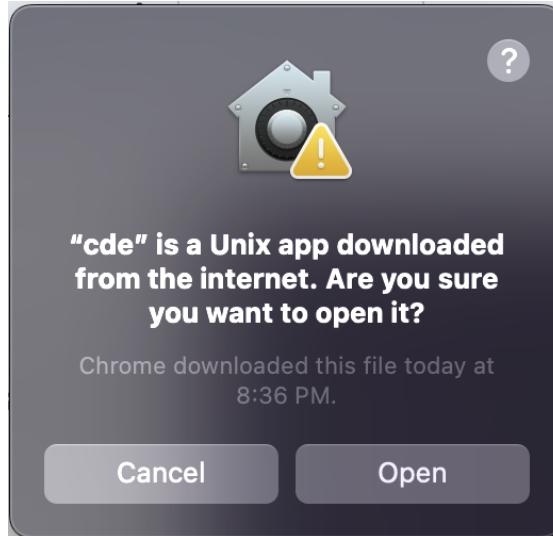
The screenshot shows the detailed view for cluster05. At the top, there's a message: "Initializing - This might take upto a couple of minutes." Below that, the cluster details are shown: VERSION 1.18.1-h3-b6, VC ID dex-app-hsvdvm7p, CREATED BY, CPU 0, and MEMORY 0 B. There are several tabs at the top: "CLI TOOL", "API DOC", "JOBS API URL", "GRAFANA CHARTS", "AIRFLOW UI", and "Events". The "CLI TOOL" tab is selected and highlighted with a red box. Below the tabs, there are three rows for different operating systems: "Mac", "Linux", and "Windows". Each row has a download icon and a file name: "Mac", "Linux", and "Windows". Underneath these, there are two tabs: "Charts" and "Logs", with "Logs" being the active tab and highlighted with a blue underline.

For Mac users:

- Make sure that the `cde` file is executable by running the below command.

```
chmod +x /path/to/cde
```

- Go to the folder where the executable is present. Right click and select “Open with” -> Terminal . You will get the below message



- Click on **Open**
- Once done, you will get the following window and message

```
--credentials-profile string    CDP credentials profile name (default "defa
ult")
  -h, --help                      help for cde
  --hide-progress-bars            hide progress bars for file uploads
  --region string                 CDP Control Plane region ("us-west-1", "eu-
1" or "ap-1") (default "us-west-1")
  --skip-credentials-file        skip CDP credentials file discovery
  --tls-ca-certs string          additional PEM-encoded CA certificates
  --tls-insecure                  skip verification of API server TLS certifi
cate
  --user string                   CDP user to authenticate as
  --vcluster-endpoint string     CDE virtual cluster endpoint
  -v, --verbose                   verbose logging
  --version                       version for cde

Use "cde [command] --help" for more information about a command.

Saving session...
...copying shared history...
...saving history...truncating history files...
...completed.

[Process completed]
```

- To validate the installation, run the below command from the terminal.
./cde --help

```
mehra@Downloads % chmod +x cde
[mehra@ Downloads % ./cde --help
Cloudera Data Engineering

Usage:
  cde [command]

Available Commands:
  airflow      Airflow commands
  backup       Create and Restore CDE backups
  credential   Manage CDE credentials
  help         Help about any command
  job          Manage CDE jobs
  resource     Manage CDE resources
  run          Manage CDE runs
  spark        Spark commands

Flags:
  --access-key-id string      access key identifier
  --access-key-secret string   access key secret
  --auth-cache-file string    token file cache location (default "$USERCACHE/token-cache")
  --auth-no-cache              do not cache authentication tokens
  --auth-pass-file string    authentication password file location
  --cdp-endpoint string       CDP API endpoint (default depends on CDP Control Plane region)
  --credentials-file string   CDP credentials file location
  --credentials-profile string CDP credentials profile name (default "default")
  -h, --help                  help for cde
  --hide-progress-bars        hide progress bars for file uploads
  --region string             CDP Control Plane region ("us-west-1", "eu-1" or "ap-1") (default "us-west-1")
  --skip-credentials-file    skip CDP credentials file discovery
  --tls-ca-certs string      additional PEM-encoded CA certificates
  --tls-insecure               skip verification of API server TLS certificate
  --user string                CDP user to authenticate as
  --vcluster-endpoint string  CDE virtual cluster endpoint
  -v, --verbose                verbose logging
  --version                   version for cde

Use "cde [command] --help" for more information about a command.
[mehra@ Downloads % ]
```

- If you get the output as shown above, then the installation is completed successfully. We now need to configure the CLI to connect to our virtual cluster.
- For configuring the CDE CLI, we create a new file and add the cluster details and use it as an environment variable for connecting to the CDE virtual cluster.
- Create a file as config.yaml and add the following details.

```
touch config.yaml
```

```
[mehra@MacBook-Pro workshop % touch config.yaml
[mehra@MacBook-Pro workshop % ls
  cde      config.yaml
[mehra@MacBook-Pro workshop % ]
```

```
vi config.yaml
```

```
user: <CDP_user>
vcluster-endpoint: <CDE_virtual_cluster_endpoint>
```

Here, **user** is the username you have been mapped in the excel sheet.

vcluster-endpoint can be obtained from the Virtual Cluster that is assigned to you. Go to the Virtual Cluster “Cluster Details”



Click on the copy icon next to JOBS API URL to copy the vcluster-endpoint

The figure shows a screenshot of the CDE Cluster Details interface for a cluster named 'cluster03'. The cluster is running. Below the cluster name, there are details: VERSION 1.18.1-h3-b6, VC ID dex-app-bcjj65n5, CREATED BY (empty), and resource usage: CPU 0, MEMORY 0 B, JOBS 0. Below these details is a navigation bar with links: CLI TOOL, API DOC, JOBS API URL (highlighted with a red box and a copy icon), GRAFANA CHARTS, and AIRFLOW UI. Underneath the navigation bar are tabs: Configuration (selected), Charts, Logs, and Access.

```
user: wuser01
vcluster-endpoint: https://hg98mkgj.cde-7fsd4m65.pse-part.dp5i-5vkq.cloudera.site/dex/api/v1
```

- Save config.yaml
- Run the below command to validate the configuration. Upon running it, you will be asked to provide the API password. Please enter the password as mentioned in the excel sheet.

```
./cde job list
```

- Once you enter the password, you should see all the jobs present in the virtual cluster.

```
mmehra@ Downloads % ./cde job list
API User Password: ----- Enter your workload password
[{
  "name": "apac52_Lab3A_access_logs_ETL",
  "type": "spark",
  "created": "2023-05-09T14:05:57Z",
  "modified": "2023-05-09T14:05:57Z",
  "retentionPolicy": "keep_indefinitely",
  "mounts": [
    {
      "resourceName": "apac52_resources"
    }
  ],
  "spark": {
    "file": "Lab3A_access_logs_ETL.py",
    "driverMemory": "1g",
    "driverCores": 1,
    "executorMemory": "1g",
    "executorCores": 1,
    "conf": {
      "dex.safariEnabled": "false",
      "spark.pyspark.python": "python3"
    },
    "logLevel": "INFO"
  },
  "schedule": {
    "enabled": false,
    "user": "host_mmehra"
  }
},
{
  "name": "apac52_Lab3B1_Data_Extraction_Sub_150k",
  "type": "spark",
  "created": "2023-05-09T14:06:40Z",
  "modified": "2023-05-09T14:29:43Z",
  "retentionPolicy": "keep_indefinitely",
  "mounts": [
    {
      "resourceName": "apac52_resources"
    }
  ],
  "spark": {
    "file": "Lab3B1_Data_Extraction_Sub_150k.py",
    "driverMemory": "1g",
    "driverCores": 1,
    "executorMemory": "1g",
    "executorCores": 1,
    "conf": {
      "dex.safariEnabled": "false",
      "spark.pyspark.python": "python3"
    },
    "logLevel": "INFO"
  },
  "schedule": {
    "enabled": false,
    "user": "host_mmehra",
    "cronExpression": "* /2 * * * *",
    "start": "2023-05-09T14:25:52.455Z",
    "end": "2023-05-10T14:25:52.455Z"
  }
},
{
  "name": "apac52_Lab3B2_Data_Extraction_Over_150k",
  "type": "spark",
  "created": "2023-05-09T14:07:10Z",
  "modified": "2023-05-09T14:36:32Z",
  "retentionPolicy": "keep_indefinitely",
}
```

- If you get any error related to the certificate, please add the flag to skip tls verification.

```
./cde job list --tls-insecure
```

- This marks the end of installation and configuration of CDE CLI. Now, head over to the next lab to trigger the jobs from CLI.

For Windows users:

- Open Powershell and navigate to the folder where you have downloaded the cde.exe file.
- You can use the below command to navigate.

```
cd C:\Users\<path-to-cde.exe folder>
```

- Run the below command to start the cde cli. It will be executed in the background.

```
start .\cde.exe
```

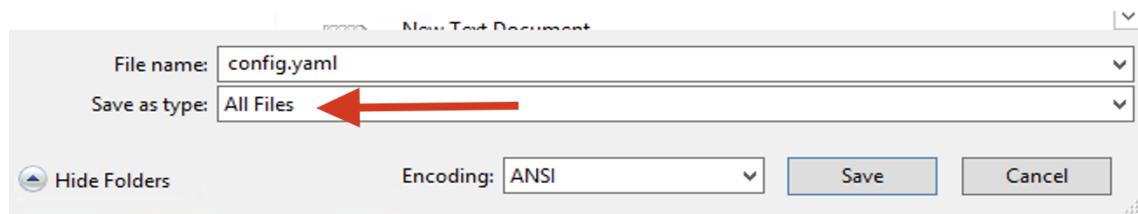
```
PS C:\Users\cdp_windows> cd C:\Users\cdp_windows\Desktop
PS C:\Users\cdp_windows\Desktop> ls

    Directory: C:\Users\cdp_windows\Desktop

Mode                LastWriteTime      Length Name
----                -----
-a---        10/28/2021  8:37 AM     38056448 cde.exe
-a---        10/28/2021  8:52 AM       106 config.yaml
-a---        10/28/2021  9:00 AM         0 he.txt
-a---        10/28/2021  8:51 AM         0 New Text Document.txt

PS C:\Users\cdp_windows\Desktop> start .\cde.exe
PS C:\Users\cdp_windows\Desktop> _
```

- Create a new text file and name it as *config.yaml*. Please note that while saving, choose the format as **All Files** and NOT as **Text Documents**.



- Add the following lines in this file.

```
user: <CDP_user>
vcluster-endpoint: <CDE_virtual_cluster_endpoint>
```

Here, **user** is the username you have been mapped in the excel sheet. For the **vcluster-endpoint** get in touch with the instructor.

[Can be obtained from your virtual cluster]

- Open Powershell and run the below command to create an environment variable.

```
$env:CDE_CONFIG = "C:\Users\<path-to-config.yaml>"
```

- Run the below command for validation. You should see the path-to-config.yaml as the output.

```
ls $env:CDE_CONFIG
```

```
PS C:\Users\cdp_windows\Desktop> start .\cde.exe
PS C:\Users\cdp_windows\Desktop> $env:CDE_CONFIG = "C:\Users\cdp_windows\Desktop\config.yaml"
PS C:\Users\cdp_windows\Desktop> ls env:CDE_CONFIG
```

Name	Value
CDE_CONFIG	C:\Users\cdp_windows\Desktop\config.yaml

```
PS C:\Users\cdp_windows\Desktop>
```

- Run the below command to validate the configuration. Upon running it, you will be asked to provide the API password. Please enter the workload password as mentioned in the excel sheet.

```
.\cde job list
```

```
PS C:\Users\cdp_windows\Desktop> .\cde job list
API User Password: _
```

- If you get the below error related to certificate, please follow the next step to skip tls verification.

```
x509: certificate signed by unknown authority
[Errno 1] _ssl.c:501: error:1409008B:SSL routines:SSL3_GET_SERVER_CERTIFICATE:certificate verify failed
```

- Run the below command with the tls flag and enter the API password.

```
.\cde job list --tls-insecure
```

```
PS C:\Users\cdp_windows\Desktop> .\cde job list --tls-insecure
WARN: Plaintext or insecure TLS connection requested, take care before continuing. Continue? yes/no [no]: yes
API User Password: _
```

- Once you enter the password, you should see all the jobs present in the virtual cluster.
- This marks the end of installation and configuration of CDE CLI. Now, head over to the next lab to trigger the jobs from CLI.

Lab 6 - Run jobs using CDE CLI

You can use the CLI to create and update jobs, view job details, manage job resources, run jobs, and so on. Please use the link below to read more about the usage of CLI to manage CDE jobs.

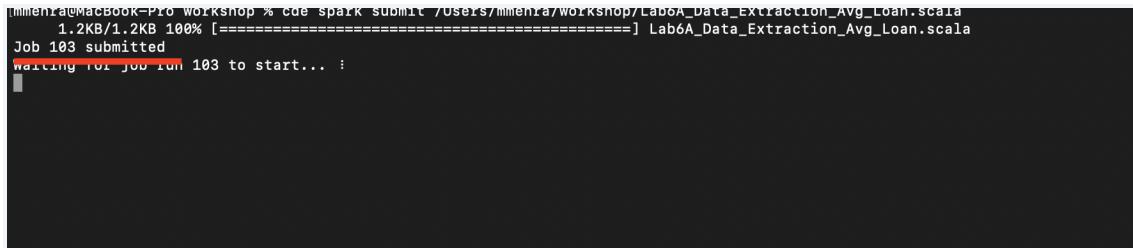
<https://docs.cloudera.com/data-engineering/cloud/cli-access/topics/cde-cli-manage-jobs.html>

Run a spark-scala job using CLI

As a first exercise in this lab, we will trigger a spark-scala job using the CDE CLI. Please note that you don't have to build a jar to submit the job to CDE.

- Locate and get the path of the script *Lab6A_Data_Extraction_Avg_Loan.scala* downloaded from the prerequisites step.
- Run the below command to submit this job to CDE.

```
./cde spark submit  
/path/to/Lab6A_Data_Extraction_Avg_Loan.scala
```



A terminal window showing the command being run and its output. The command is `./cde spark submit /path/to/Lab6A_Data_Extraction_Avg_Loan.scala`. The output shows the file being uploaded (1.2KB/1.2KB 100%), the job being submitted (Job 103 submitted), and the job status (Waiting for job run 103 to start...).

- Go to CDE UI and click on Job Runs. You will see a job submitted with the name `cli-submit-<username>-<temp-resource-id>`

Run ID	Job ↑	Type	User	Duration
103	cli-submit-wuser01-1635928294681	Spark	wuser01	1.3 MIN

- You can observe the logs and SparkUI for this Job Run.
- Please note that you are not creating this as a job in CDE. It will be an ad-hoc run without the need of registering it as a job.

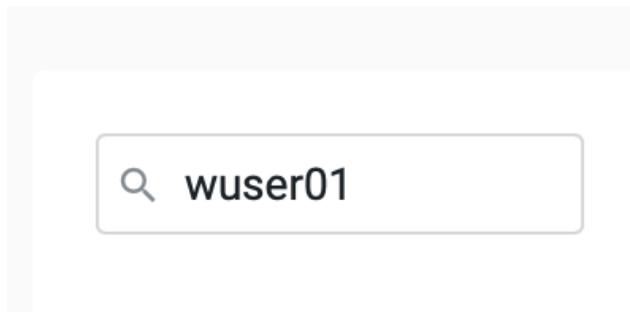
Lab 7 - Data Lineage and Auto-Scaling

In this lab, you will go through the data lineage of the two use cases that we worked on. Additionally, you will also see the auto-scaling capabilities of CDE service with the rising demand for compute resources.

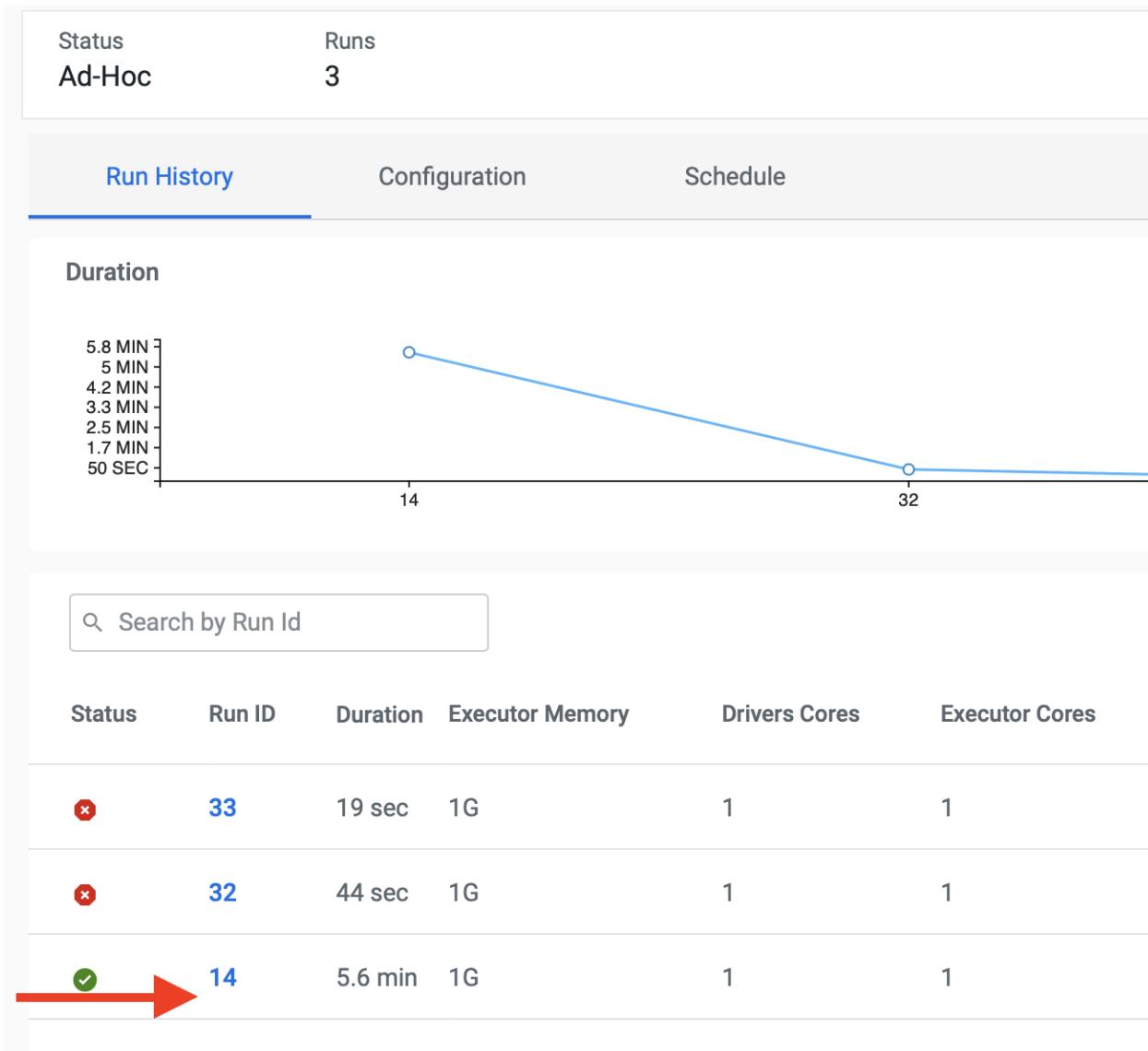
Data Lineage using Atlas

- In the CDE UI, click on the Jobs tab. Go to the job `<username>_Lab3B3_Create_Report`s that you have created in the Lab2.
- To get the jobs, please filter the jobs with your username.

Jobs



- In Run History tab, click on the successful Run ID i.e., the one with the green tick mark.



- Click on **Atlas** under Lineage.

Job Runs / 14

The screenshot shows a UI component with two sections. On the left, under 'Status', there is a green checkmark icon followed by the word 'Success'. A large red arrow points from this section to the right. On the right, under 'Lineage', the word 'Atlas' is displayed in blue text.

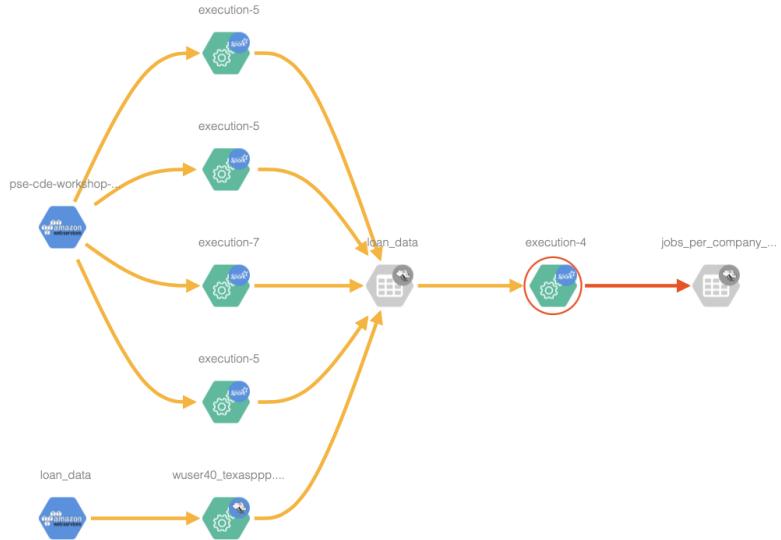
- Click on the execution that you see in the list.

Name	Owner	Description	Type	Classifications	Term
execution-4			spark_process	+	+

- Click on **Lineage** to observe the Data Lineage for this job.

The screenshot shows a navigation bar with several tabs: 'Properties', 'Lineage' (which is highlighted in green), 'Relationships', 'Classifications', and 'Audits'. To the left of the tabs, the word 'Terms:' is followed by a green square button containing a white plus sign.

○ Current Entity ⏳ In Progress → Lineage → Impact

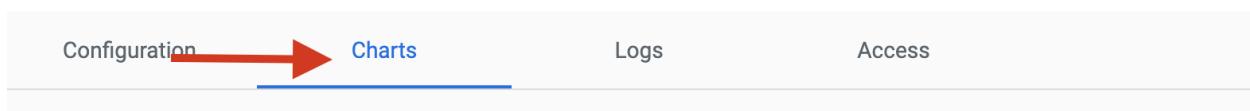


- Click on each entity to understand how the data is flowing from source to consumption.

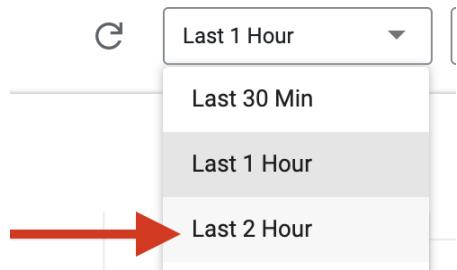
Auto-scaling in CDE

- As a last step, we want you to witness the auto-scaling capabilities of CDE. At the start of the lab, you might have noticed the cpu and memory consumption of the virtual cluster. Please check the dashboard now to see how it has scaled up based on the demand experienced.
- On the CDE home page, click on the **Cluster Details** on the virtual cluster.
- Click on the **Charts** tab.

CLI TOOL : API DOC JOBS API URL | GRAFANA CHARTS | AIRFLOW UI



- Set the filter to **Last 2 Hour** and observe the varying load on cpu and memory.



- Click on **Grafana Charts** to view another set of metrics of the virtual cluster.

CLI TOOL : [API DOC](#) [JOBS API URL](#) **GRAFANA CHARTS** [AIRFLOW UI](#)

- This marks the end of the overall CDE Hands-on Workshop session.

THANK YOU VERY MUCH FOR YOUR PARTICIPATION