Data Lifecycle on CDP Public Cloud

Data Engineering Lab

Goals:

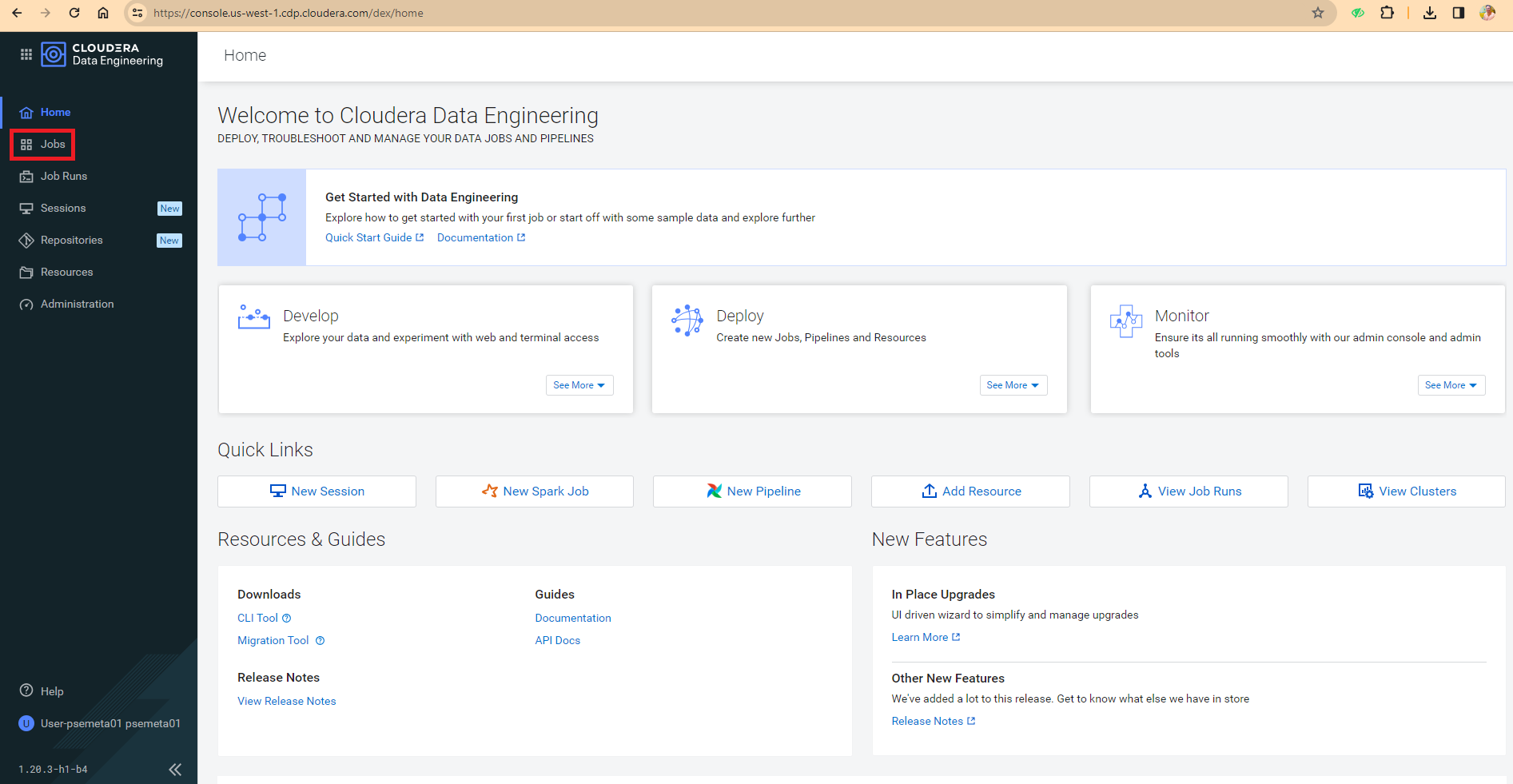
* Run a data enrichment process
* Run a process to simulate changes to the data
* Configure the execution of a pipeline using low-code/no-code tools

1. Click on DataFlow from CDP PC Home:

A screenshot of a computer

Description automatically generated

1. The Data Engineering Home shows all the actions that can be done, such as Jobs in Spark and pipelines in Airflow, Resources and useful information/documentation. Click on the option **Jobs** from the left menu to create a dataflow in Airflow.



1. Here the available tasks are listed. For the purposes of this workshop, two Jobs have been configured:
   * **\_CDE-Table-Update**, generate random changes and enrich table to visualize Lakehouse Time Travel functionality.
   * **\_CDE-Data-Enrichment**, process in Spark (Python) to enrich the data ingested from Kafka and save to a new table.

It is time to create our Job in Airflow. Click on **Create Job.**

A screenshot of a computer

Description automatically generated

1. In the Job creation form, you must enter the following information:

* Job Type: **Airflow**
* Name: Use the naming <assigned user>-pipeline. Replace <assigned user> with the user assigned to you. For example, ***psemeta01***
* DAG***:* Editor*,*** to graphically configure the task. Then, click **Create**.

A screenshot of a computer

Description automatically generated

1. On the Job editing screen, select the Editor tab, and you will see the following canvas to drag the steps of the pipeline that we are going to create. In our case, we are going to create two CDE Jobs and relate them. Drage the **CDE job** into the canvas.

A screenshot of a computer

Description automatically generated

1. Let's start with the first Job. Click on the CDE Job button and drag onto the canvas, entering the following settings:

* **Select Job**: select the Job ***\_CDE-Data-Enrichment***
* Check the checkbox **Override Spark values**. Additional options will appear below.
* **Arguments**: <assigned user>. Use the username assigned to you. For example, ***psemeta01***

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Configure the second Job. Click on the CDE Job button and drag onto the canvas, entering the following settings:

* **Select Job**: select the Job ***\_CDE-Table-Update***
* Check the checkbox **Override Spark values**. Additional options will appear below.
* **Arguments**: <assigned user>. Use the username assigned to you. For example, ***psemeta01***

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. To set up the execution sequence, bind **cde\_job\_1** with **cde\_job\_2**. For that, click on the right connector of the job of **cde\_job\_1** and drag to the left connector of **cde\_job\_2**.

A screenshot of a computer

Description automatically generated

Once the Jobs are linked let’s rename the jobs. Click on **cde\_job\_1** and then rename it as **Data Enrichment.**

A screen shot of a computer

Description automatically generated

Click on **cde\_job\_2** and then rename it as **Table Update.**

A screenshot of a computer

Description automatically generated

1. Once the Jobs have been joined, click on **Save** to save the settings made. You should see a message indicating **Pipeline saved to job**.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. The time has come to run the pipeline. On the upper right side of the canvas, click **Actions**

**-> Run Now**.

A screenshot of a computer

Description automatically generated

1. You should see the pipeline execution screen, indicating that the execution has been initialized.

A screen shot of a computer

Description automatically generated

Also, on the Job Runs tab you can see the pipeline and the very first job of the pipeline getting started.

A screenshot of a computer

Description automatically generated

After some time the second job starts and then the pipeline and the two jobs are completed.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Click on the Airflow UI tab to see the execution detail of each step in the pipeline. The configured Data Enrichment and Table Update jobs are listed at the bottom left. The colours indicate the status of each job. Make sure the radio button **Auto-refresh** is enabled to automatically display the status of jobs.

A screenshot of a computer

Description automatically generated

1. You can see more information about the execution by clicking on the view **Graph**. Hovering the mouse over the Job name displays specific information for each step in the pipeline. Make sure the pipeline status is Success, which indicates that the entire pipeline was able to run without issue.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

*The execution status appears next to the name of the pipeline (marked in red). If it is green and indicates* ***Success****, it means that the execution was successful.*

A screenshot of a computer

Description automatically generated