Plan the Environment

You will need to deploy the following services in CDP PC:

* Light Duty Data Lake
  + Do not use any automation tool to deploy the environment, as you cannot get fully public endpoints using them. Use the manual steps described in the [AWS Quickstart](https://docs.cloudera.com/cdp-public-cloud/cloud/aws-quickstart/topics/mc-aws-quickstart.html#mc-aws-quickstart) guide. Make sure the enable “Public endpoints” during the setup process.
* One Data Hub
  + Streams Manager (Kafka) [template](https://docs.cloudera.com/data-hub/cloud/overview/topics/dh-streams-messaging-clusters.html)
* CDF Data Service
  + Minimum of 3 nodes, Recommended 5
  + Maximum as the number of users you will get in the session (safe range)
* CDE Data Service
  + Minimum of 1 node
  + Maximum as the number of users you will get in the session (safe range)
  + Only one Virtual Cluster is required, with 100% of the resources configured for the service. **Enable Spark 3 and Iceberg analytics tables**
* CDW Data Service
  + Group of 10 attendees per Impala Virtual Warehouse (this is for attendees)
  + Group of 10 attendees per Data Viz
  + One Hive Virtual Warehouse (just to run the create tables)
* CML Data Service
  + Only one Workspace is required.
  + Minimum of 3 nodes, Recommended 5
  + Maximum as the number of users you will get in the session (safe range)

# 

# Environment Checklist

Make sure to perform all this checklist before you execute the workshop labs.

The current English manuals for the labs can be found [here](https://drive.google.com/drive/folders/1va6fWTv9iCbW3-1LhzX1UDF0Em_E-YrI?usp=drive_link).

Other Details

A screenshot of a computer

Description automatically generated

## Permissions:

* Permissions in Ranger

Add the user group to the following policies:

**Hadoop SQL:**

all - database, table, column

all - storage-type, storage-url

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Kafka (Data Hub cluster name):**

all - consumergroup

all - topic

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Idbroker

**Add the user group and instructors users to the Idbroker mapping**

<https://docs.cloudera.com/cdf-datahub/7.2.9/nifi-hive-ingest/topics/cdf-datahub-hive-ingest-idbroker-mapping.html>

A screenshot of a computer

Description automatically generated

* Environment permission level:

Grant the following permissions to the user group to which all users belong to.

A screenshot of a computer

Description automatically generated

* Group permission level:

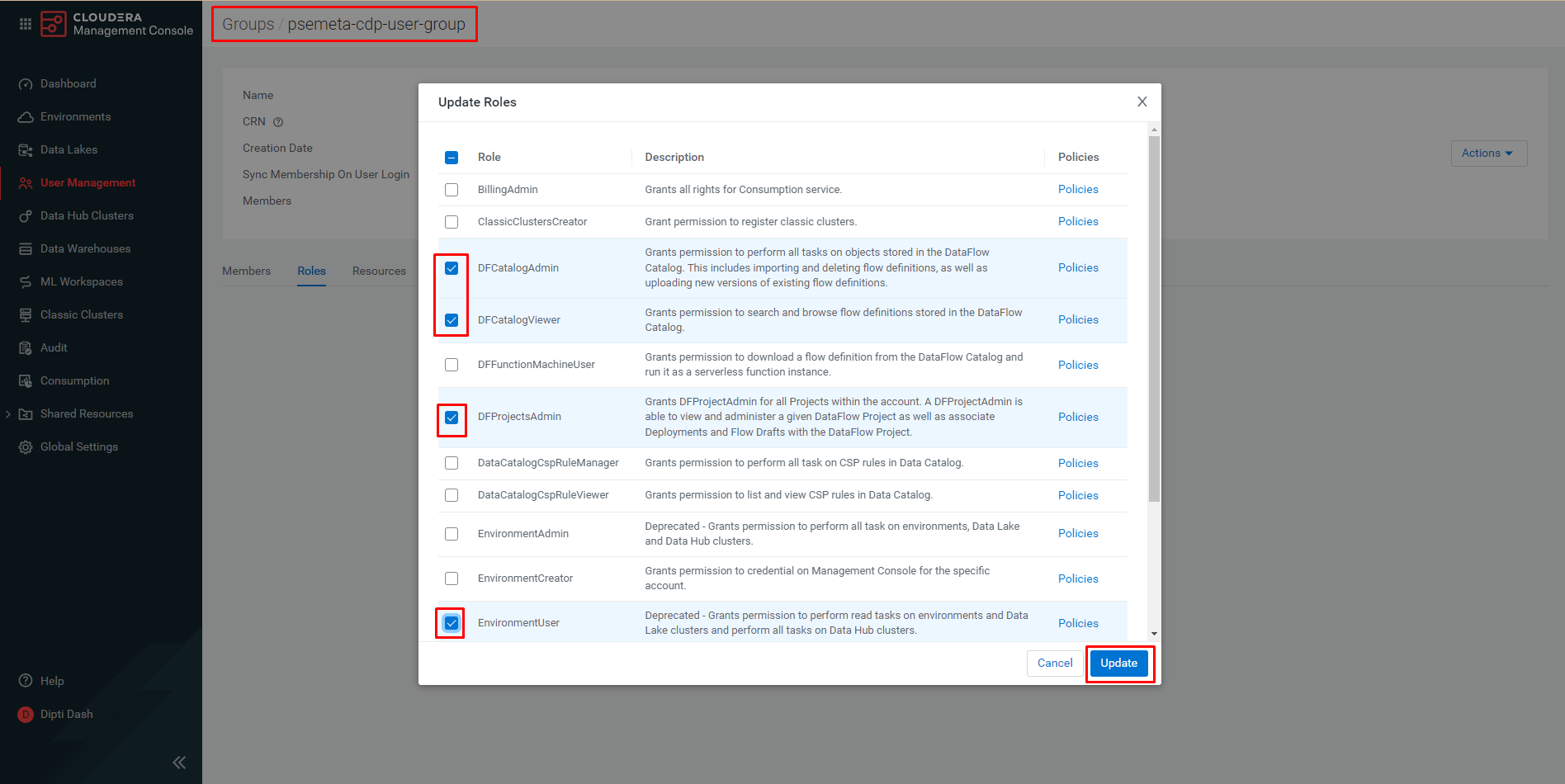
Grant the following permissions to the user group to which all users belong to

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

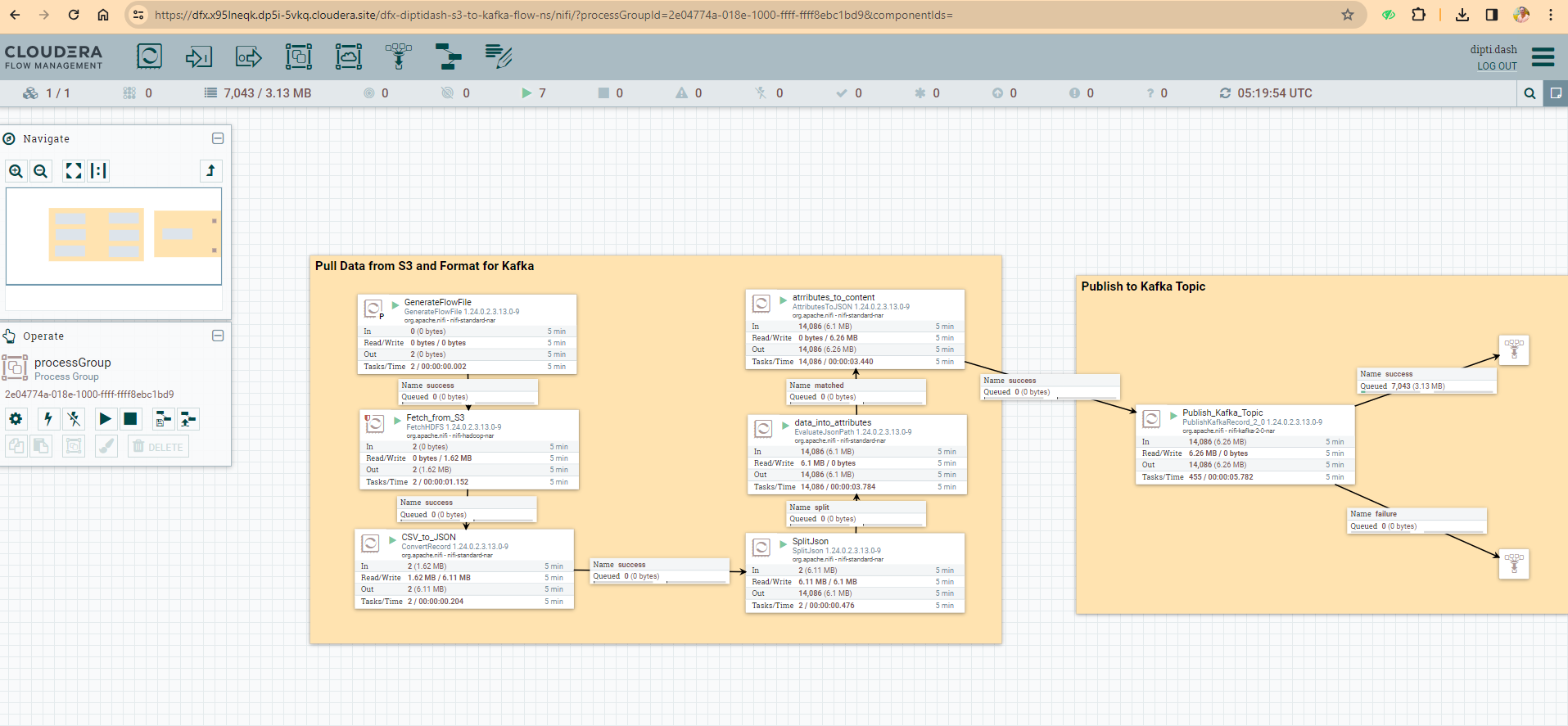


## Data Hub:

* Brand new **telco\_data** Kafka topic, no consumer groups registered. It doesn’t need to be created upfront.
* **Note**: If you have run this workshop the day before, delete the Kafka topic (use SMM for that) and load the data again using the CDF flow definition.

## Data Flow:

* [**kafka\_to\_lakehouse**](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/kafka_to_lakehouse.json) flow definition loaded in the Catalog (**Name**: lab\_kafka\_to\_lakehouse)
* [**s3\_to\_kafka**](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/s3_to_kafka.json) flow definition loaded in the Catalog (**Name**: lab\_s3\_to\_kafka)
* Execute **s3\_to\_kafka** to load data from s3 file to **telco\_data** Kafka topic
* Delete **s3\_to\_kafka** flow definition loaded from the Catalog (you dont need to show this to attendee)



A screenshot of a computer

Description automatically generated

## Data Warehouse:

* Create as many Impala VW you need for your attendee group. Recommendation: 1 VW per group of 10 people.
* Create as many Data Viz you need for your attendee group. Recommendation: 1 Data Viz per group of 10 people
* Configure a Data connection from Data Viz to Impala VW. Data Viz 1 to VW 1, Data Viz 2 to VW 2, and so on.
* Configure *URLs that remote data columns can be fetched from* foreach Data Viz to reach CML model url/workspace (**Note**: This can be fetched only after you have deployed the model as admin)
* Create one Hive VW, then run following:
  + Create or recreate databases and tables per each user ([link](https://raw.githubusercontent.com/campossalex/ClouderaHandsOnDL/main/resources/hive_drop_create_tables.txt))

A screenshot of a computer

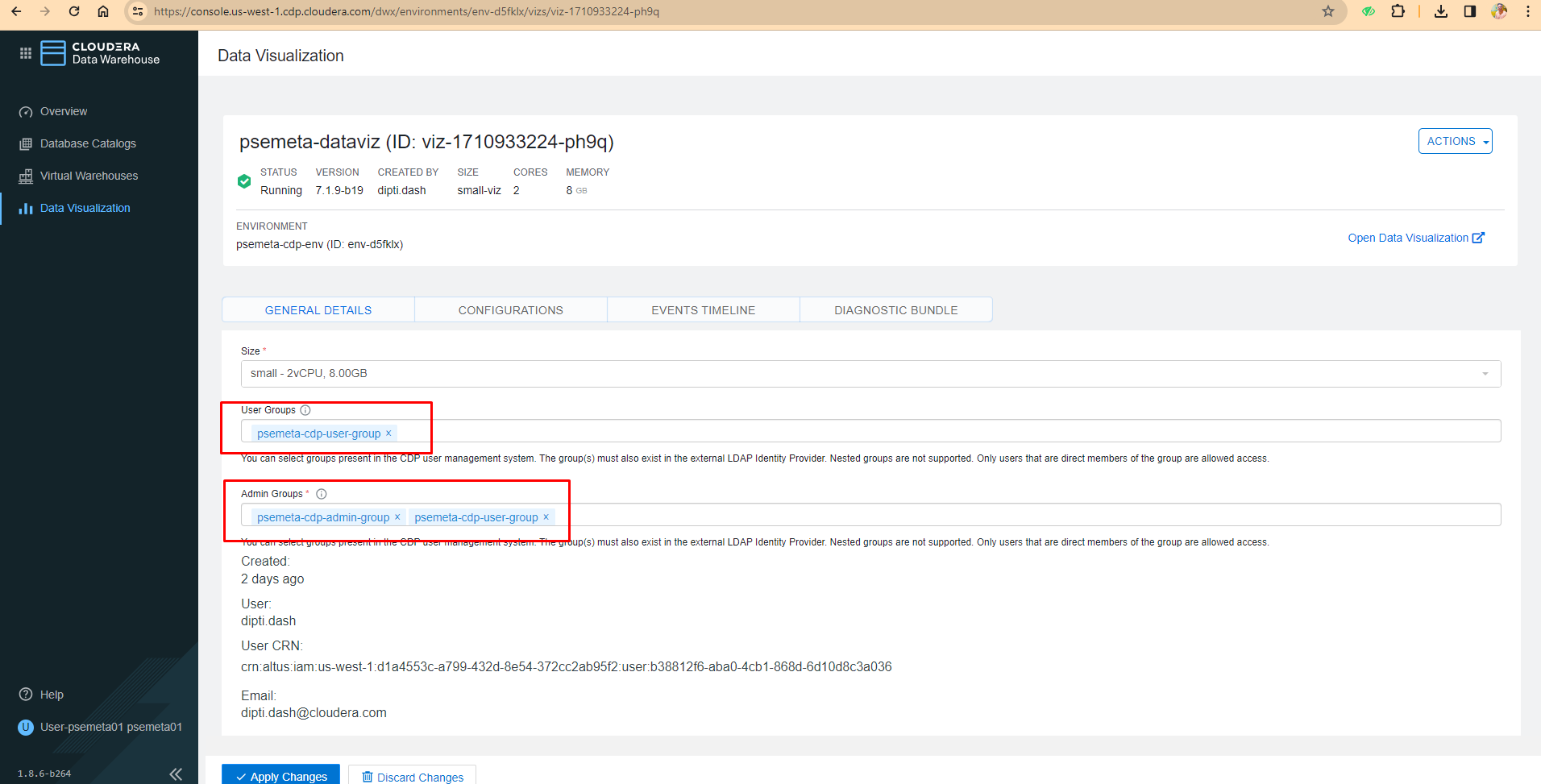
Description automatically generated

* + Create the master tables for CDE lab ([link](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/lab_guide.md#3-create-master-tables))

A screenshot of a computer

Description automatically generated

In Data Viz make sure the following permission is done for ‘New Connection’ to appear.



## Data Engineering:

* Set your environment bucket name (edit line 2) in [parameters.conf](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/parameters.conf) file.
* Upload [**parameters.conf**](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/parameters.conf), [**CDE-telco\_data\_enrichment.py**](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/CDE-telco_data_enrichment.py) and [**CDE-updateTable.py**](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/CDE-updateTable.py) scripts to a **Telco\_Churn** folder in Resources.
* Create Spark jobs for each script, without **Arguments** (do not specify the user id). Set the job name as the same as the python script (it will be easier to identify). Just **Create**, not **Create and Run.**

## Machine Learning:

* Create a workspace.
* Copy the workspace url and set a new url to add to Data Viz Remote Data Setting, e.g.: <https://modelservice>.<workspace\_url>/model

# Tips

# Dont stop the environment

Regions to pick up: regions that support K8S (EKS)

Environment creation:

* AWS Quickstart
* Automation tool

# 

# Environment Setup steps (WIP)

Setup a CDP PC environment, setting up the following services:

## Light Duty Data Lake

## Data Hub, Streams Messaging Light Duty

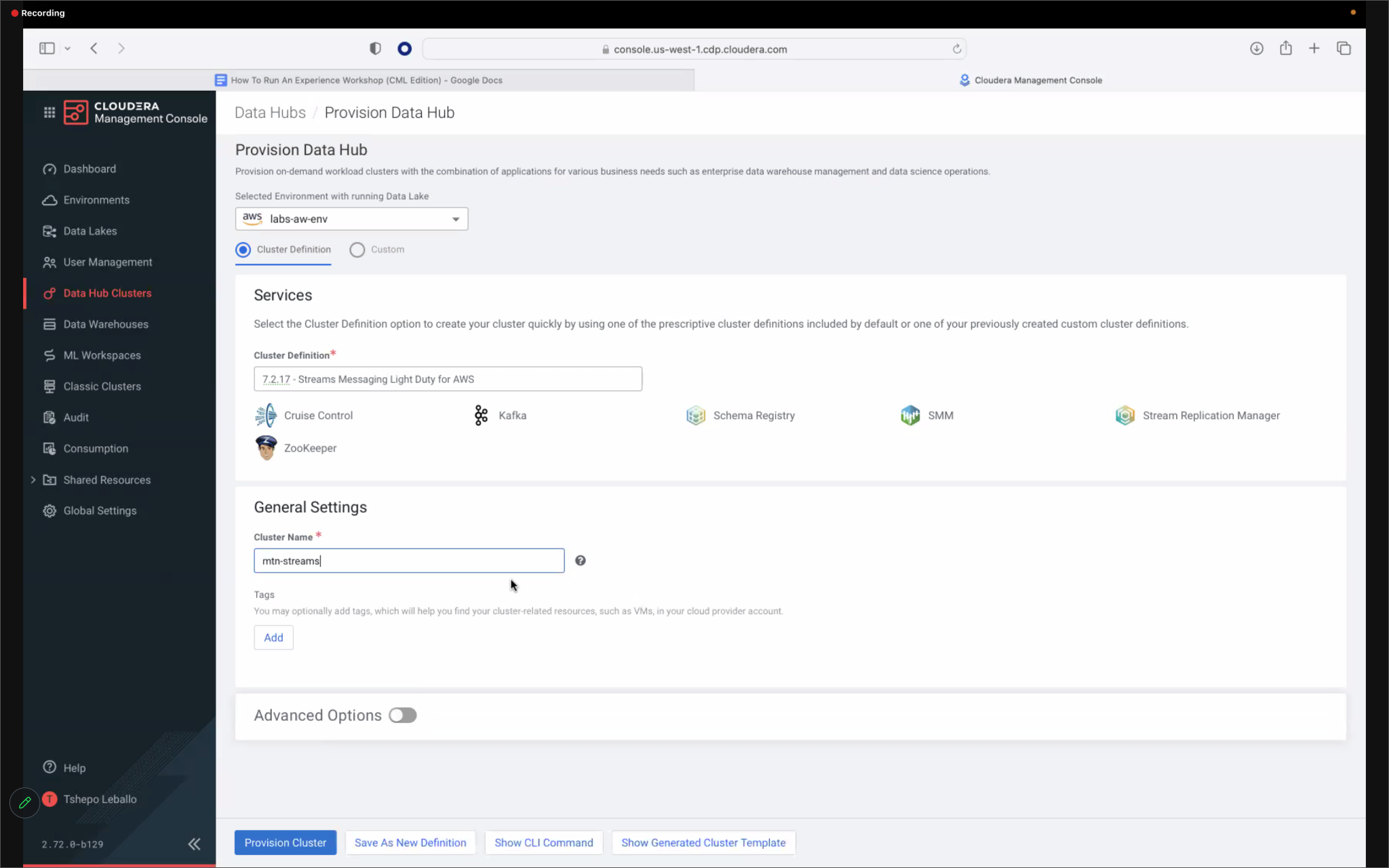
It will be used to publish and consume the main data source for the use case

### Setup Service:

Data Hub setup steps [here](https://docs.cloudera.com/data-hub/cloud/create-cluster-aws/topics/mc-create-cluster-from-template.html)

Note: Make sure to select **Streams Messaging Light Duty** cluster definition

Setup Data Hub as below:



Copy the Kafka brokers URL list. Once the Data Hub is deployed, click on it to see the information. Go to **Endpoints** -> **Kafka Broker Endpoint(s)**, and copy that value. You will need it when publishing the data to the Kafka topic (**s3\_to\_kafka**), also delivering the session when the attendees deploy the **kafka\_to\_iceberg** Flow Definition

## Data Flow

Consume data from Kafka topic and store an Iceberg table

### Service setup:

CDF setup steps [here](https://docs.cloudera.com/cdp-public-cloud-patterns/cloud/bi-at-scale/topics/pat-bias-enable-df-environment.html)

Setup CDF as below:



### Upload the data:

In the s3 bucket your environment uses as main storage, create a data directory and upload this [file](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/WA_Fn-UseC_-Telco-Customer-Churn-new.csv) to that folder. You should have something like this:

*s3a://environment-bucket/data/WA\_Fn-UseC\_-Telco-Customer-Churn-new.csv*

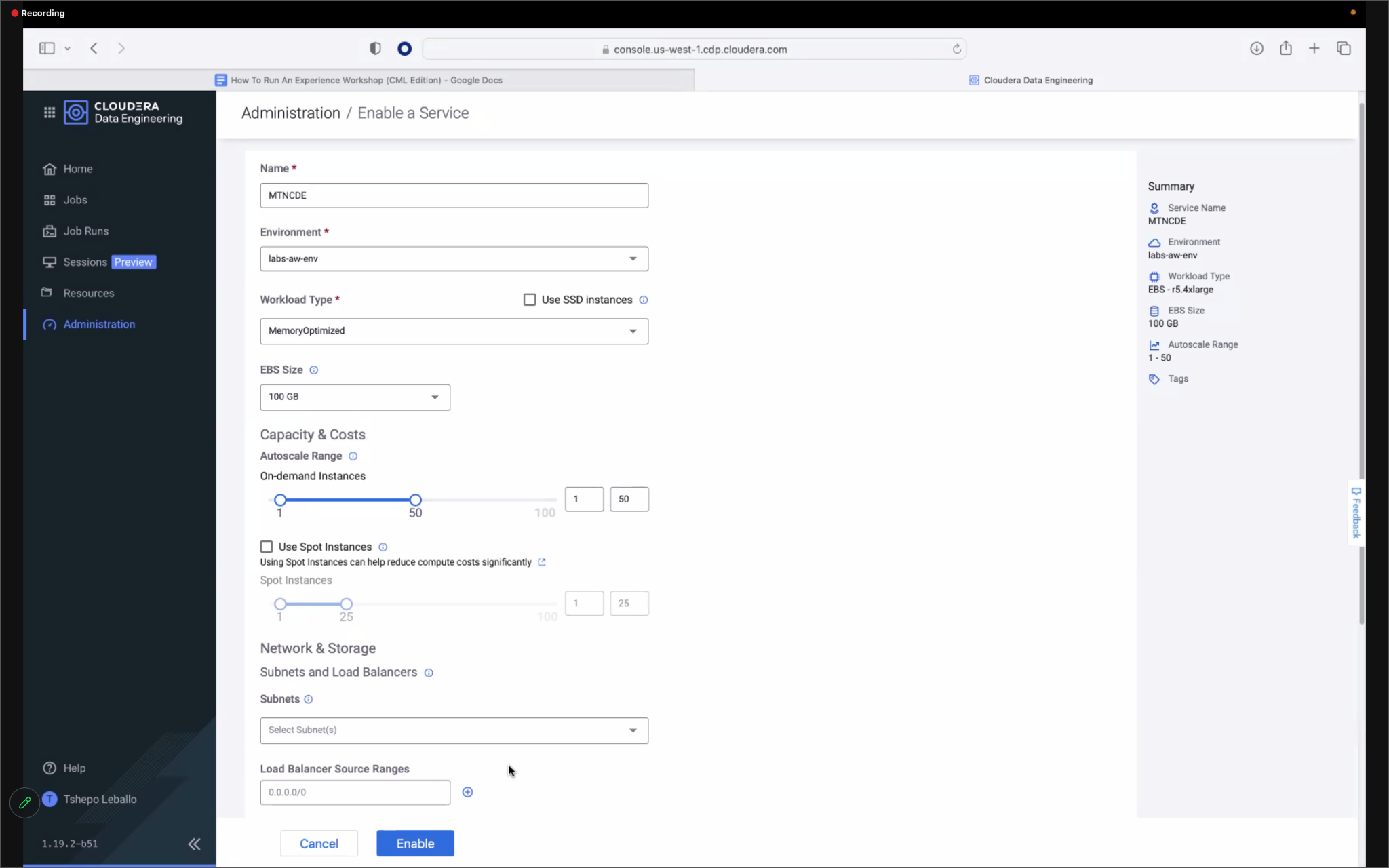
Then use the [s3\_to\_kafka](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/s3_to_kafka.json) Flow Definition to read the data from the s3 bucket and publish it in a Kafka topic. This is the first contact the attendees have with the dataset.

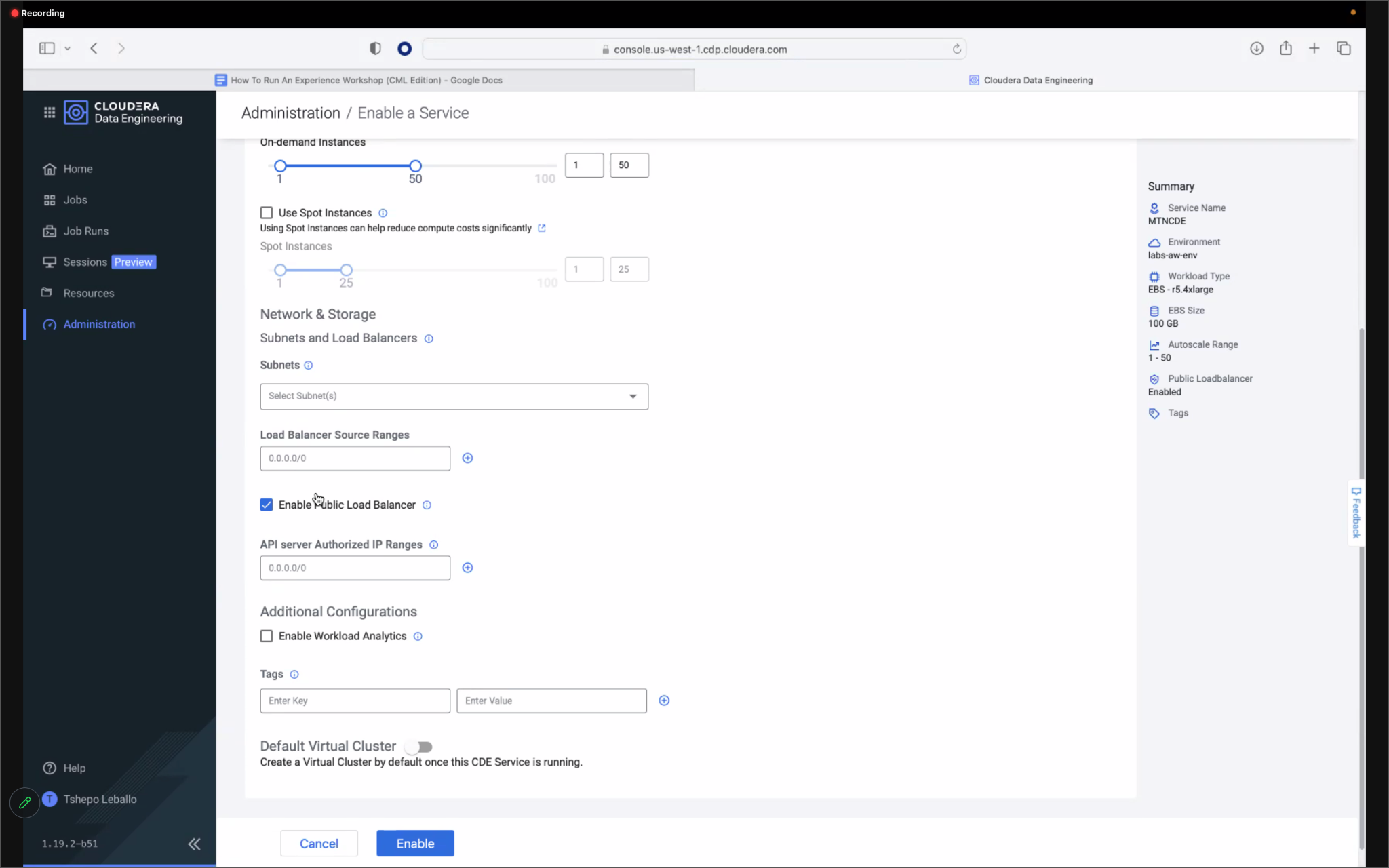
* Upload [s3\_to\_kafka](https://github.com/campossalex/ClouderaHandsOnDL/blob/main/resources/s3_to_kafka.json) Flow Definition to CDF Catalog
* Deploy the s3\_to\_kafka Flow Definition, setting the right parameters to read the data from the csv file stored in the s3 bucket and publish in the Kafka topic.

## Data Engineering

## 

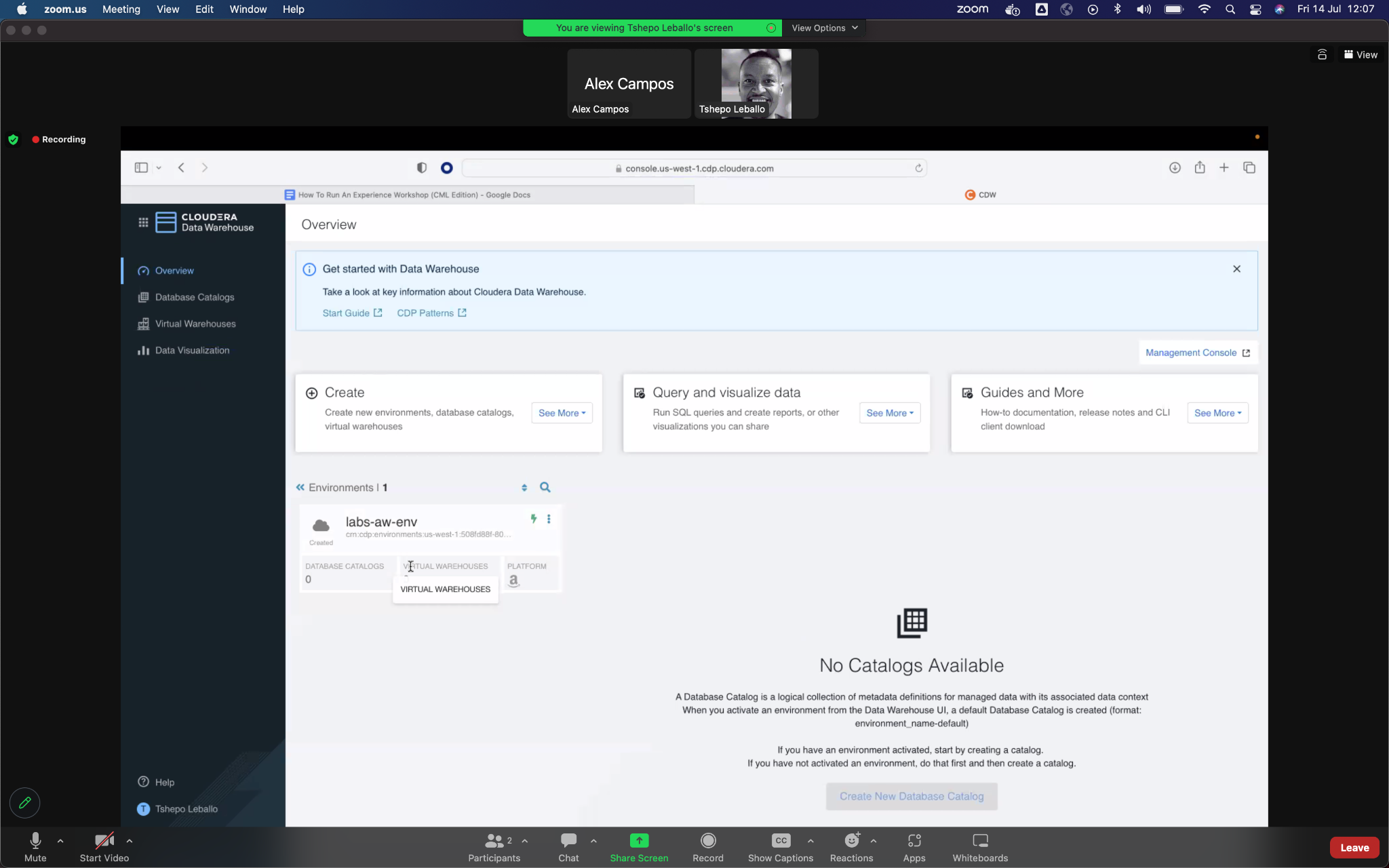
Data preparation

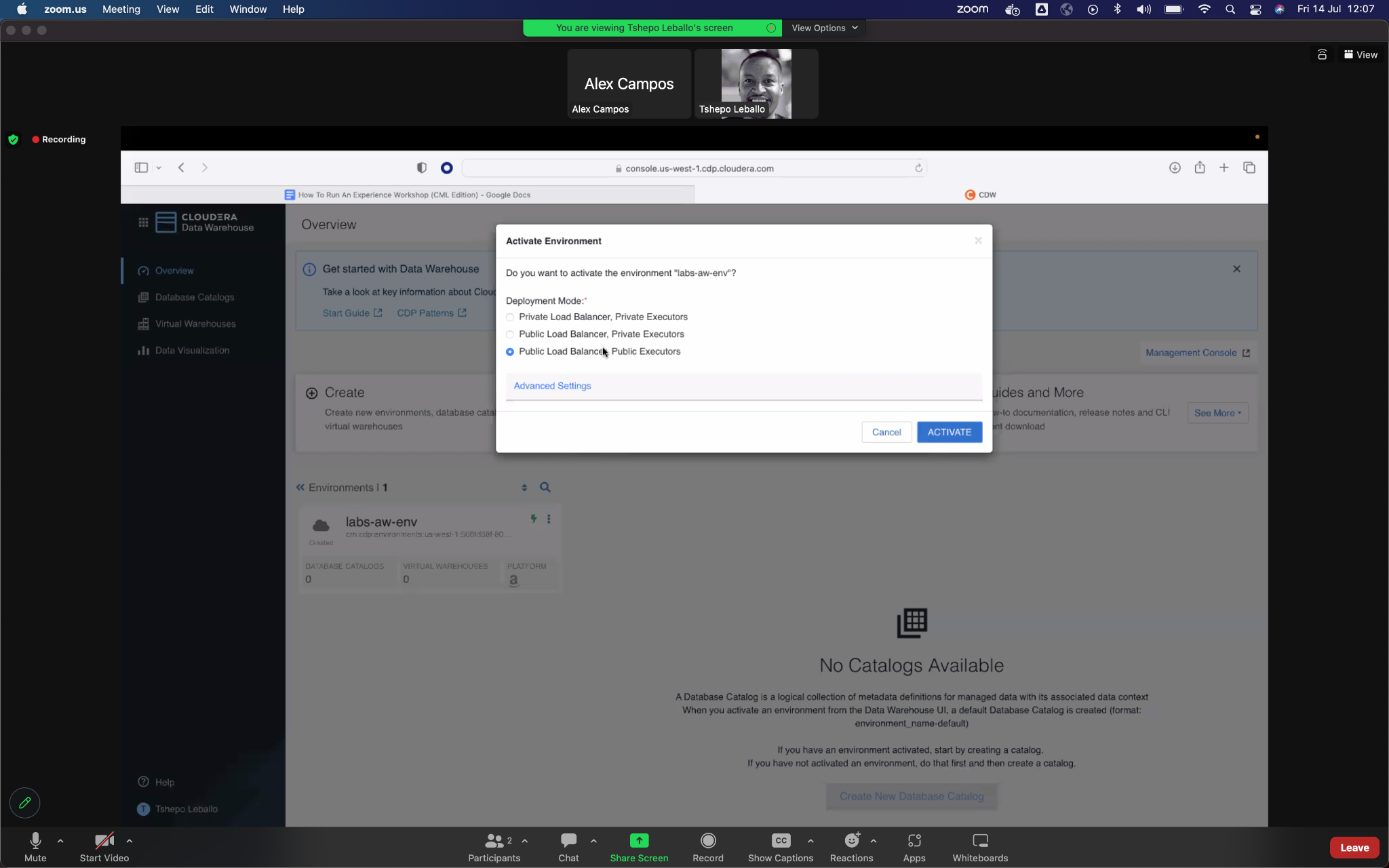




## Data Warehouse

Data exploration and build reports





## Machine Learning

Training a ML model to score customer churn

