Case Study

on

**IBM DataStage**

**Case study by,**

Mouli S,

Batch – 02,

moulisankar2002@outlook.com

**Contents**

1. Understandings of Case Study ……………………………………………3
2. Case Study Questions ………………………………………………………….4-13
   1. Question 1 …………………………………………………………4
   2. Question 2 …………………………………………………………9

**1. Understandings of Case Study**

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-COV-2 virus.

Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illness. Anyone can get sick with COVID-19 and become seriously ill or die at any age.

European Centre for Disease Prevention and Control (also known as ECDC) - an agency of the European Union tracks COVID-19 cases and vaccination status worldwide.

Reference Website - <https://www.ecdc.europa.eu/en/>

**2. Case Study Questions**

**2.1 Question 1**

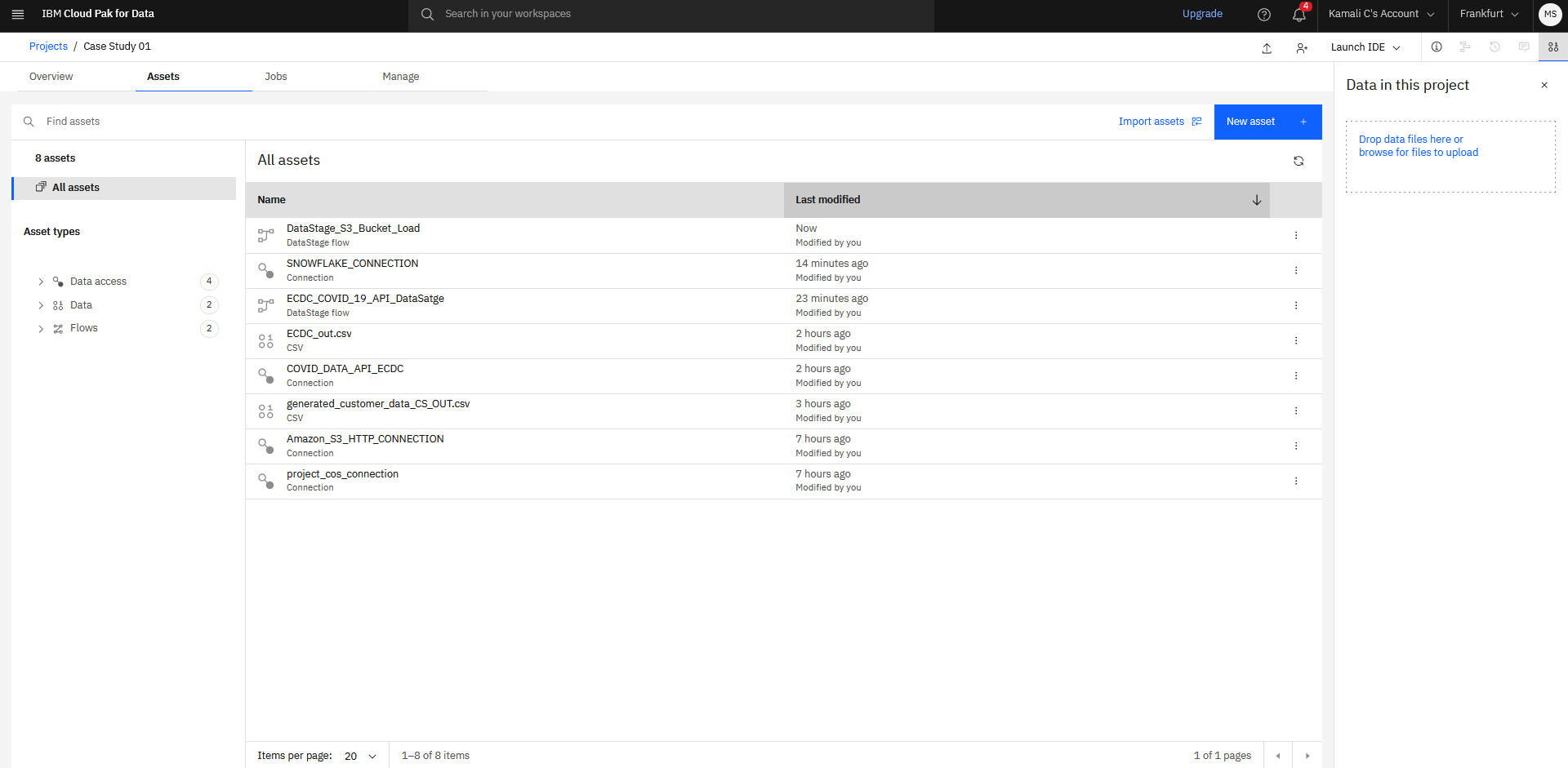
**Participants will create materialized view to copy data from S3 bucket for customer data and display out having customers whose age is greater than 30 years and less than 50 years with following columns:**

**a. CustomerName**

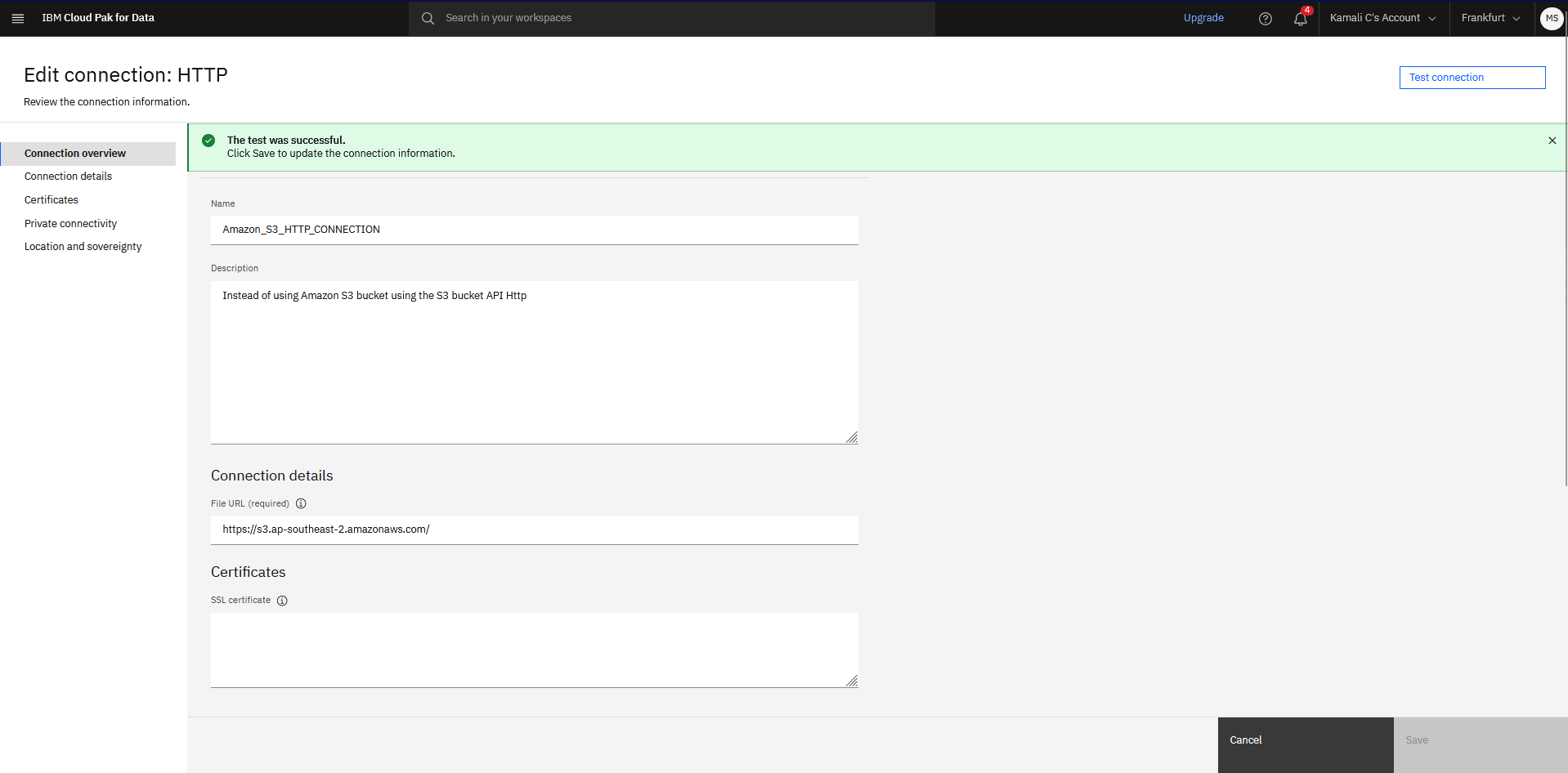
**b. CustomerAge (as on today, Integer)**

**c. CustomerCity**

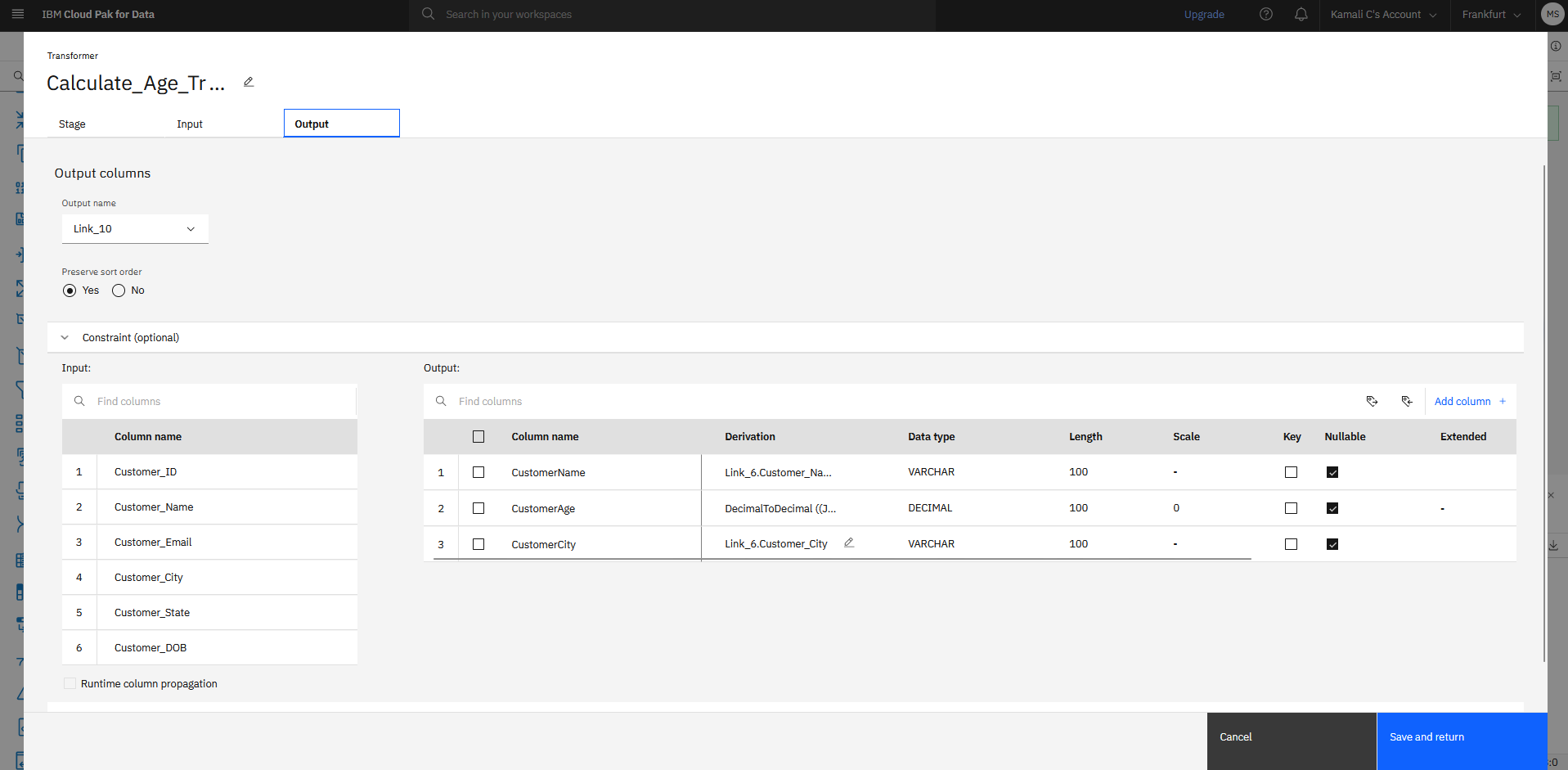
**Created Assets and Data Stages:**

****

**HTTP Connection for S3 Bucket:**

****

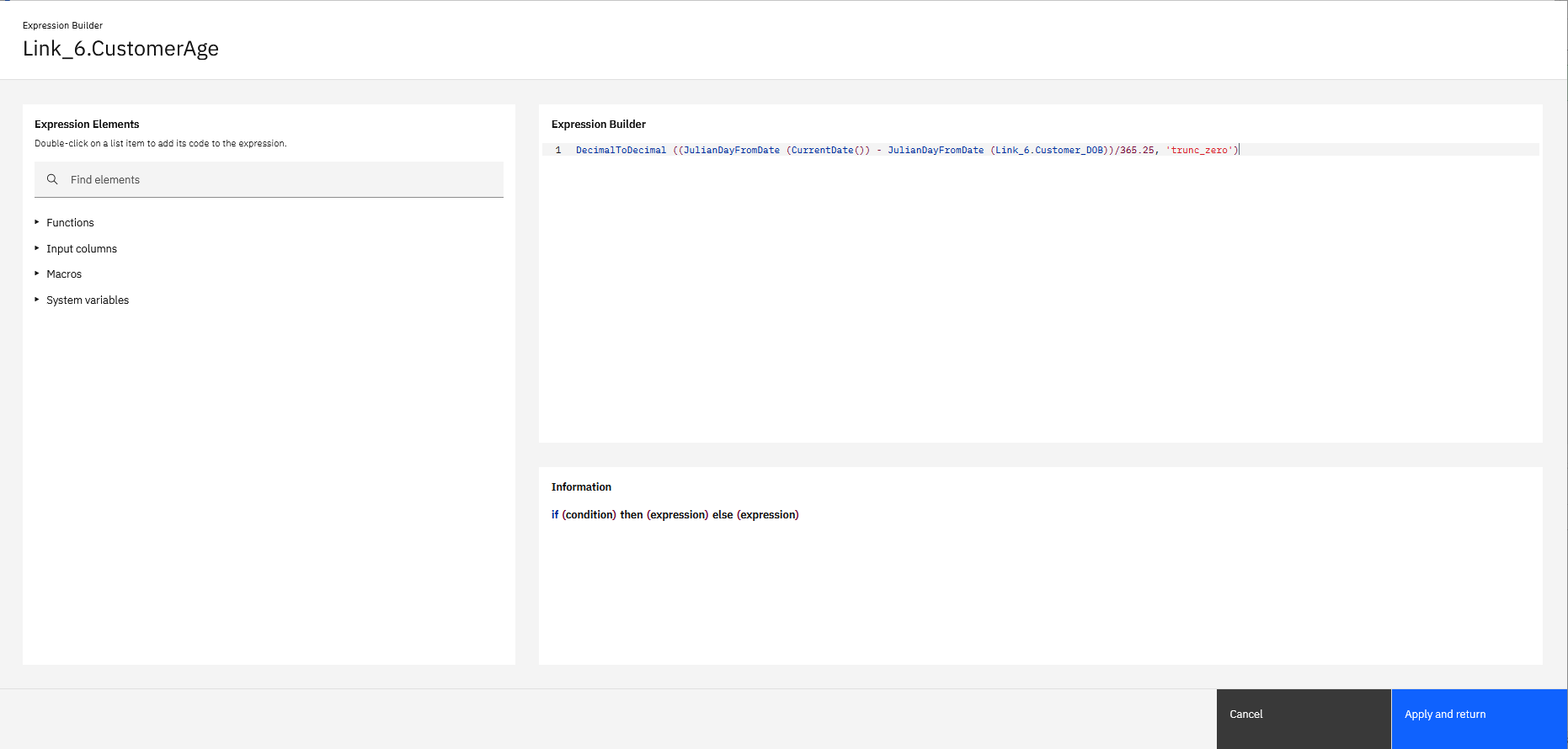
**Transformer Stage to calculate the Age of the Customers using the DOB:**

****

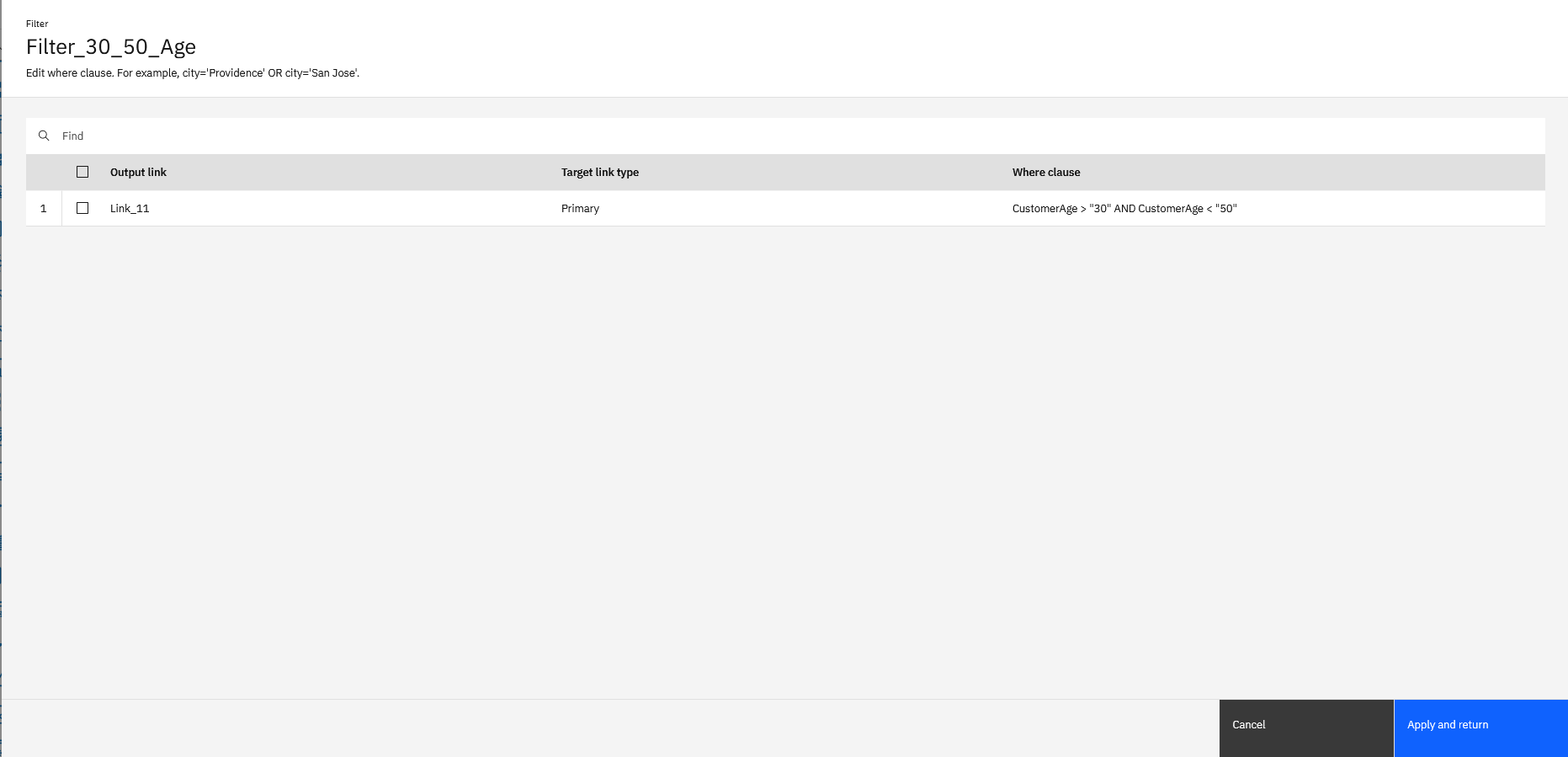
**Used Expression while calculating the Age:**

“DecimalToDecimal ((JulianDayFromDate (CurrentDate()) - JulianDayFromDate (Link\_6.Customer\_DOB))/365.25, 'trunc\_zero')”

Here I used the Julian Date function which will give the days difference in the IBM DataStage.

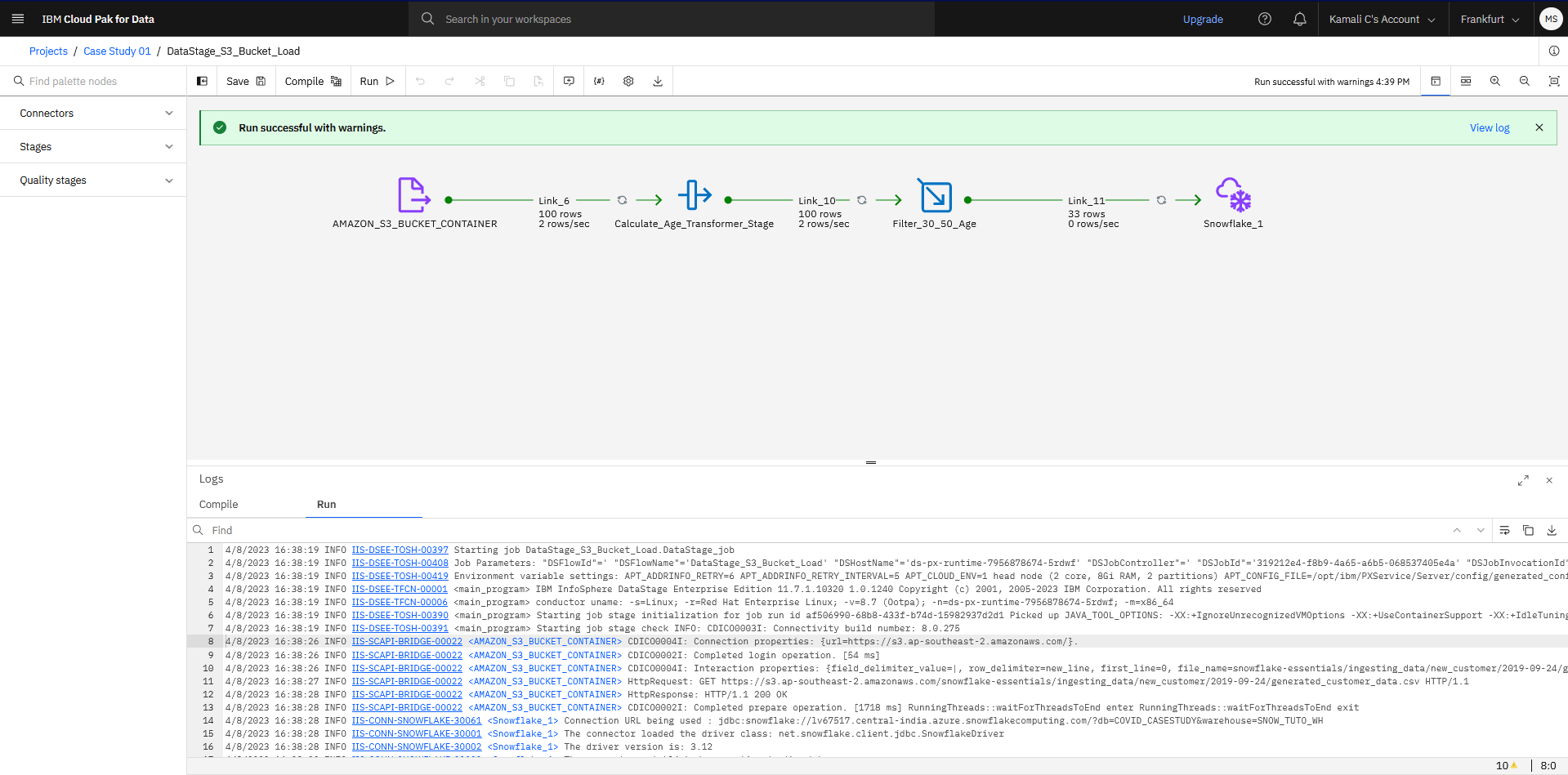
****

**Filter Stage to filter the age of customer > 30 and < 50:**

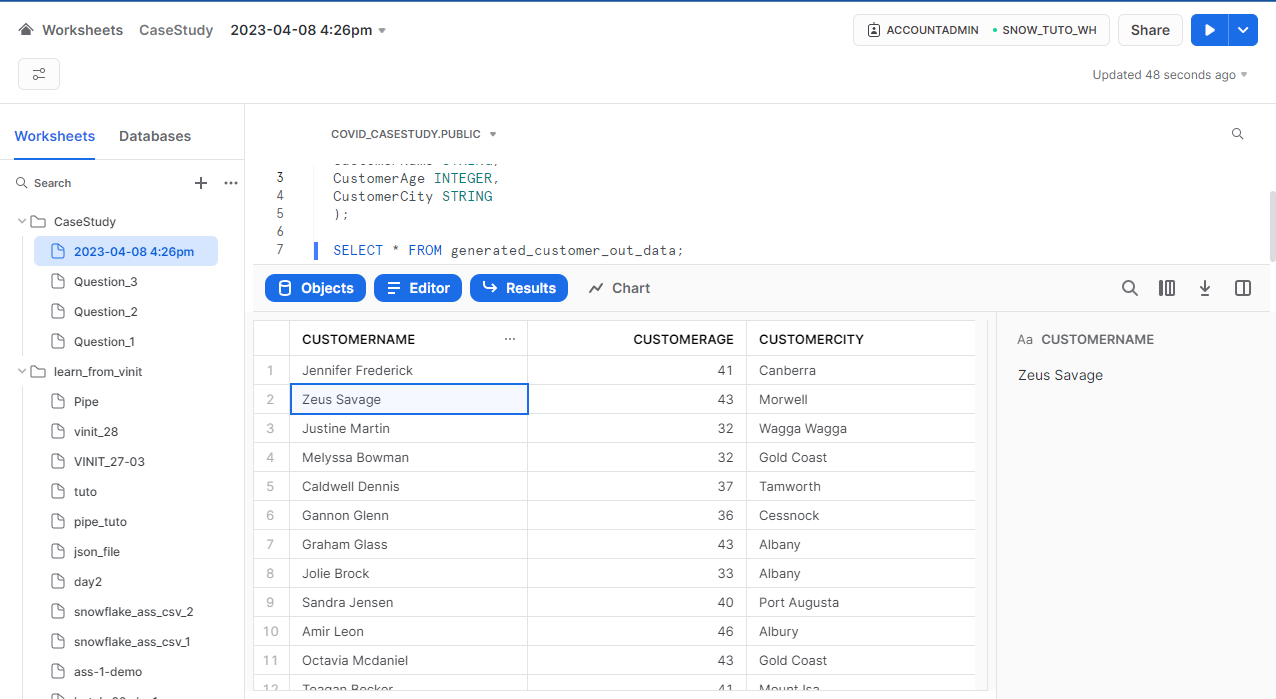
****

**Job Run:**

**Job run status. It resulted around 33 rows.**

****

**Data Preview in the Snowflake Database:**

****

**2.2 Question 2**

**Participants will use the HTTP Connection and will create the IBM DataStage job using the section 2. 2.. Output table and file will consist of following columns for Year 2020:**

**a. Continent (e.g... Asia)**

**b. Countries (e.g... Afghanistan)**

**c. Year Month (e.g., 2020-01)**

**d. Cases (e.g., 0)**

**e. Deaths (e.g., 0)**

**Data Cleansing:**

4/8/2023 19:23:48 ERROR IIS-CONN-NGBR-00035 <HTTP\_1> Error when checking operator: Error creating schema ( record( dateRep:nullable date; day:nullable int32; month:nullable int32; year:nullable int32; cases:nullable int32; deaths:nullable int32; countriesAndTerritories:nullable string[max=1024]; geoId:nullable string[max=1024]; countryterritoryCode:nullable string[max=1024]; popData2019:nullable int32; continentExp:nullable string[max=1024]; Cumulative\_number\_for\_14\_days\_of\_COVID-19\_cases\_per\_100000:nullable dfloat; ) ) from DataSetDef provided by connector: **In field "Cumulative\_number\_for\_14\_days\_of\_COVID": Expected ";", got: "-19", line 1**

4/8/2023 19:23:48 ERROR IIS-CONN-NGBR-00028 <HTTP\_1> Error when checking operator: Could not build output schema from DataSetDef provided by producer

4/8/2023 19:23:48 ERROR IIS-DSEE-TFSR-00019 <main\_program> Could not check all operators because of previous error(s)

4/8/2023 19:23:48 ERROR IIS-DSEE-TCOS-00029 <main\_program> Creation of a step finished with status = FAILED.

4/8/2023 19:23:48 INFO IIS-DSEE-TOSH-00398 Job ECDC\_COVID\_19\_API\_DataSatge.DataStage\_job aborted

So, I performed data cleansing process by using the HTTP Connection and download the data.csv file from the source in ECDC Website.

I used SSMS and SQL Server to perform cleansing process.

I change the column name from **“Cumulative\_number\_for\_14\_days\_of\_COVID-19\_cases\_per\_100000” to “Cumulative\_number\_for\_14\_days\_of\_COVID\_19\_cases\_per\_100000”**

And I exported the data as data.csv and uploaded it to my **Azure Blob Storage** in the container called **“mou-casestudy”**

**SQL Queries used for cleansing:**

-- Step 1: Add a new column with the desired name

ALTER TABLE [dbo].[data\_20230330]

ADD Cumulative\_number\_for\_14\_days\_of\_COVID\_19\_cases\_per\_100000 VARCHAR(100);

-- Step 2: Update the new column with the data from the old column

UPDATE [dbo].[data\_20230330]

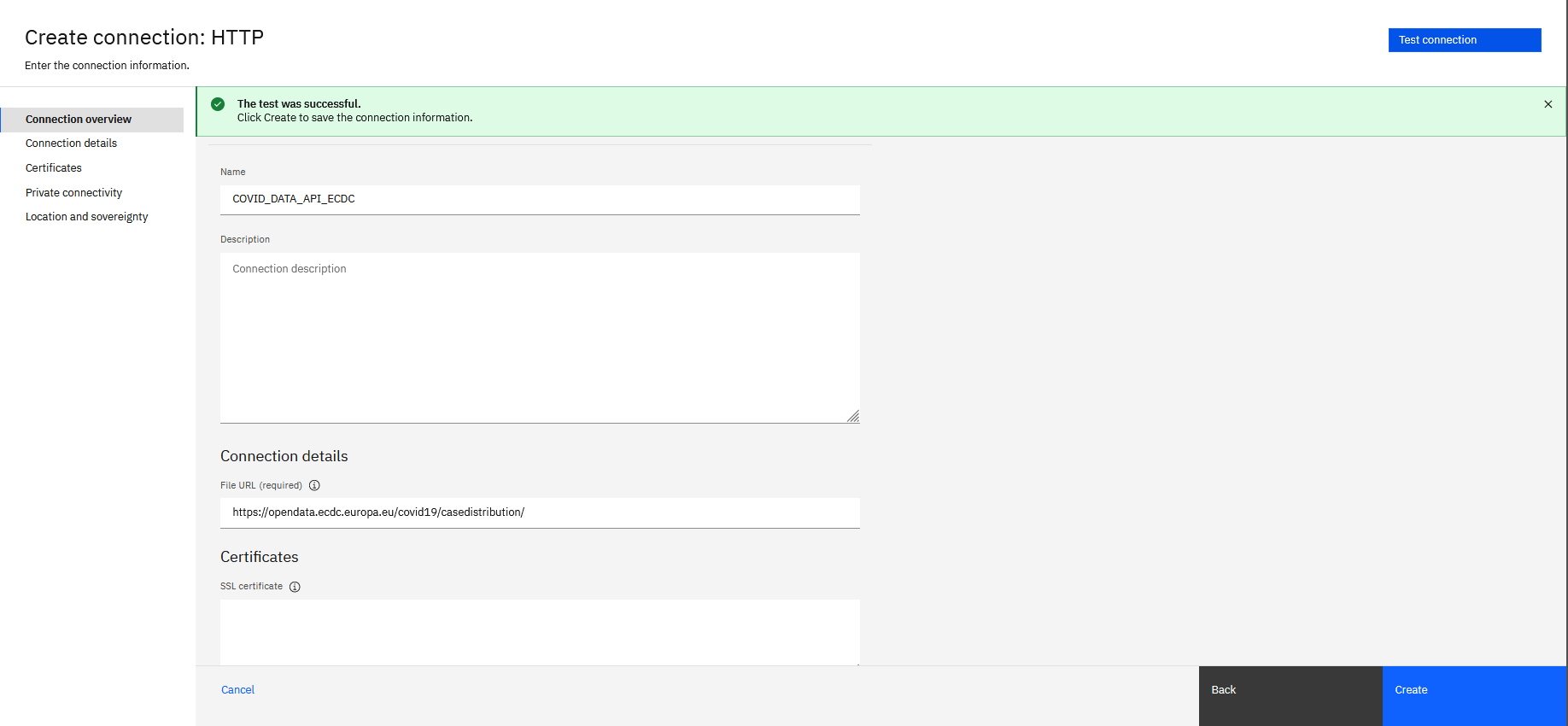
SET Cumulative\_number\_for\_14\_days\_of\_COVID\_19\_cases\_per\_100000 = [Cumulative\_number\_for\_14\_days\_of\_COVID-19\_cases\_per\_100000];

-- Step 3: Drop the old column

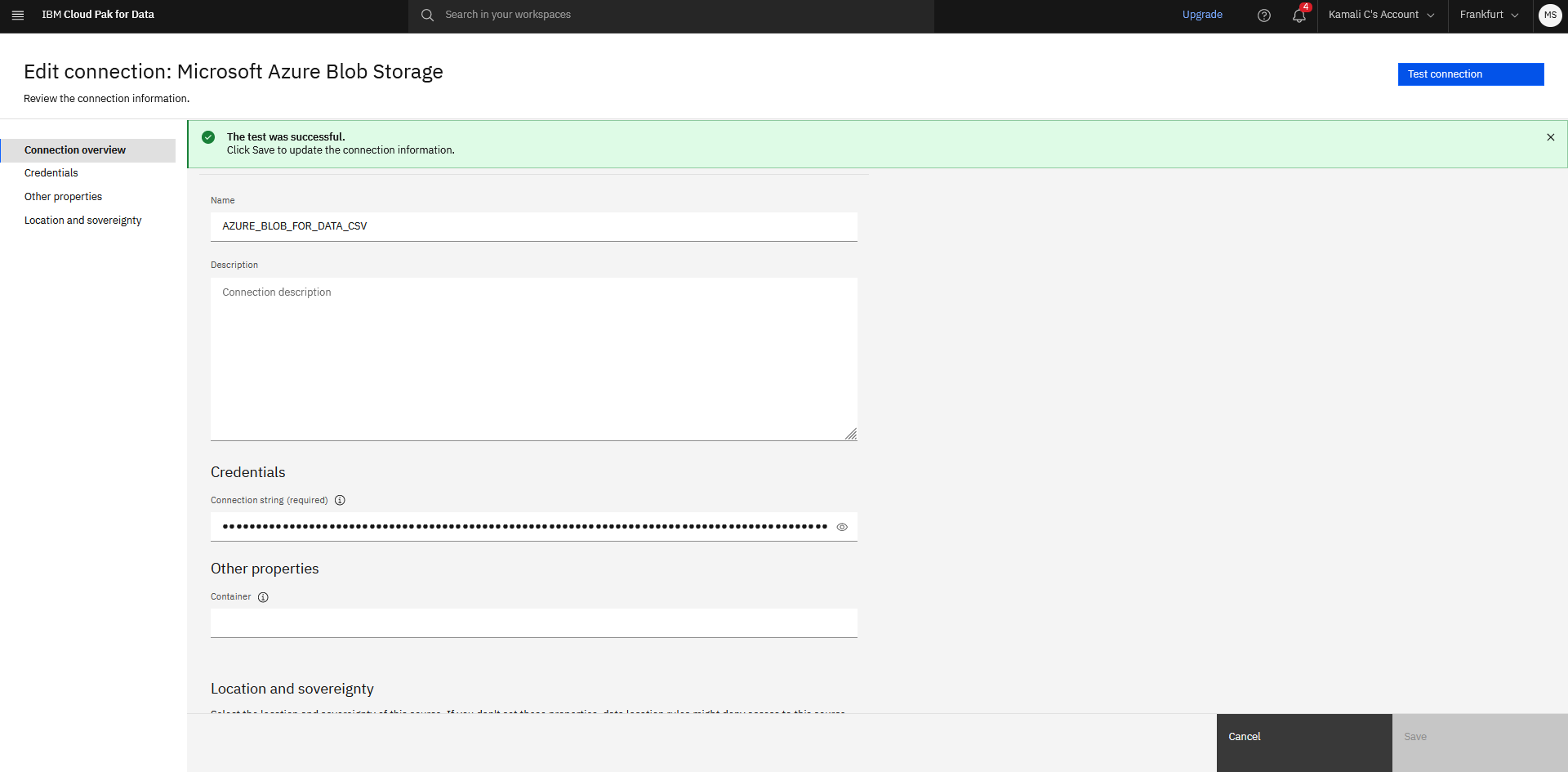
ALTER TABLE [dbo].[data\_20230330]

DROP COLUMN [Cumulative\_number\_for\_14\_days\_of\_COVID-19\_cases\_per\_100000];

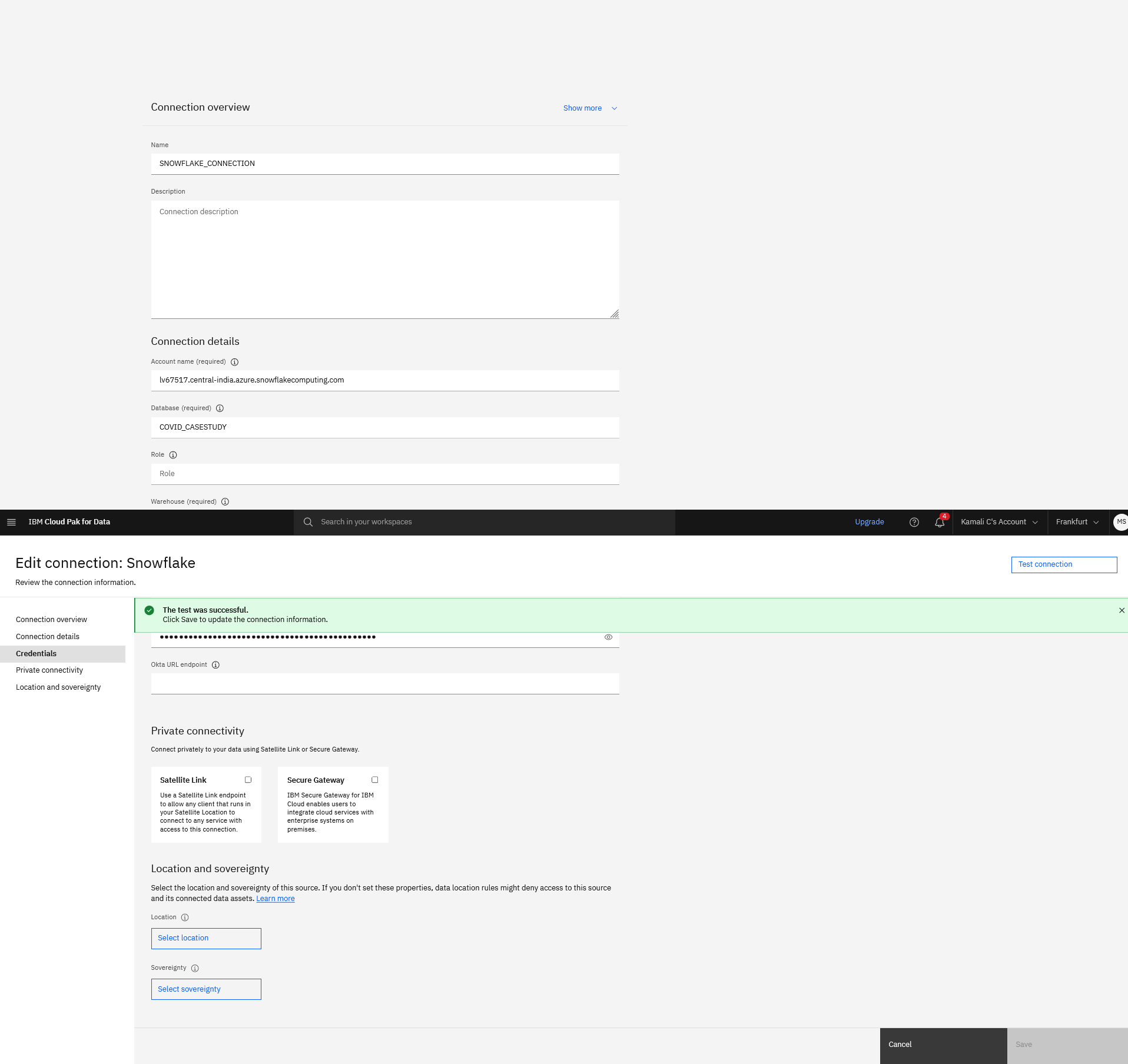
**HTTP Connection to download the data from ECDC Website:**

****

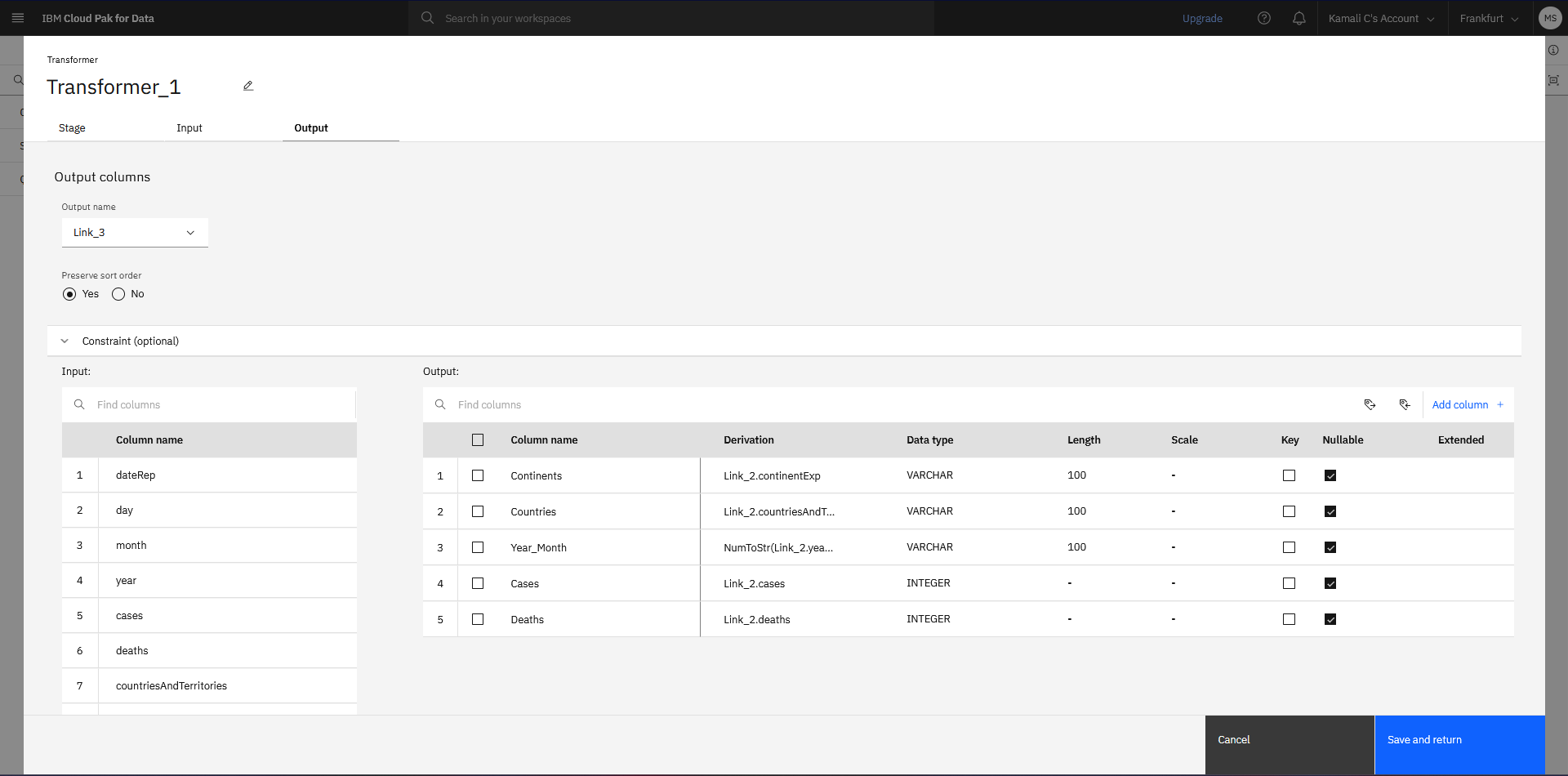
**Azure Blob Connection:**

****

**Snowflake Connection:**

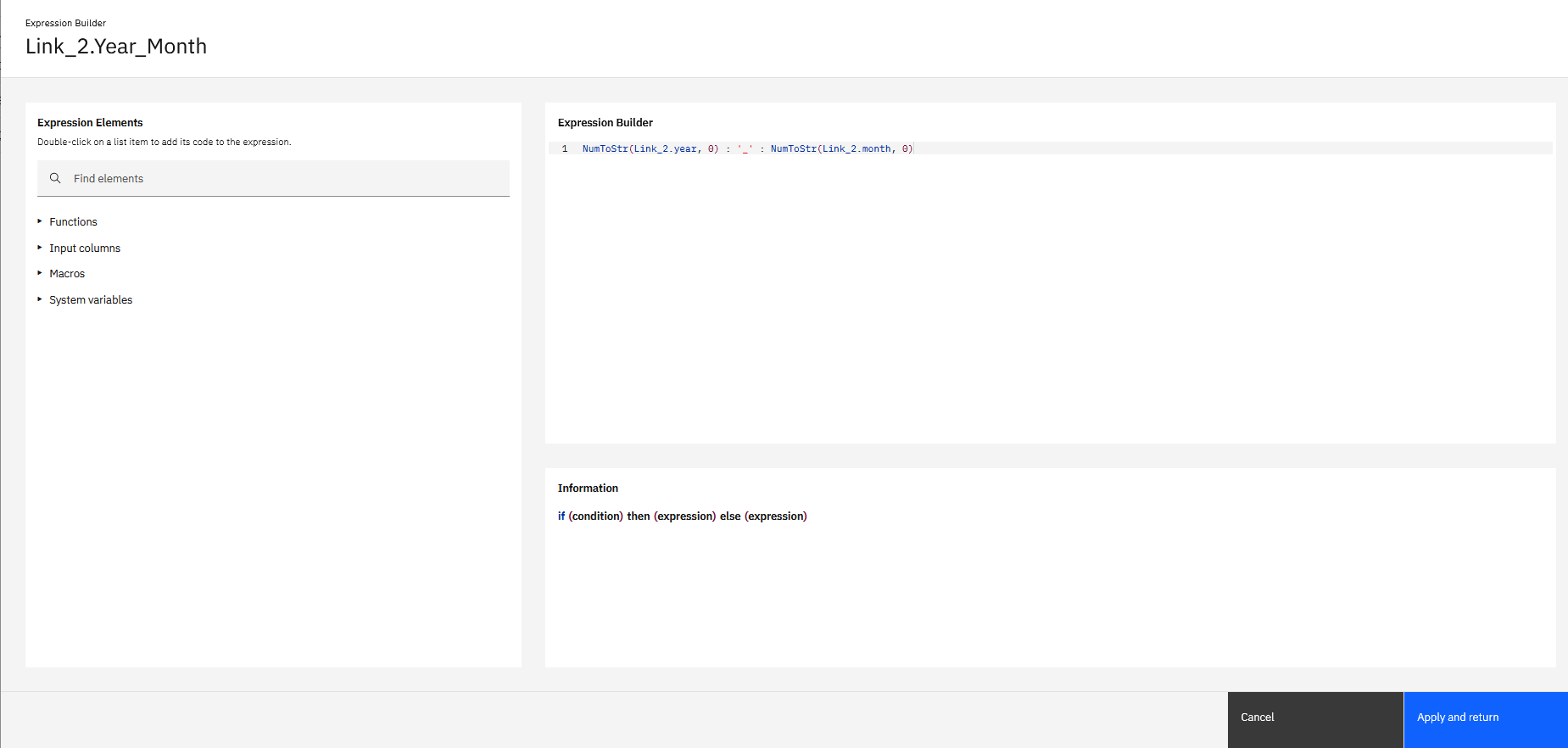


**Transformer Stage:**

****

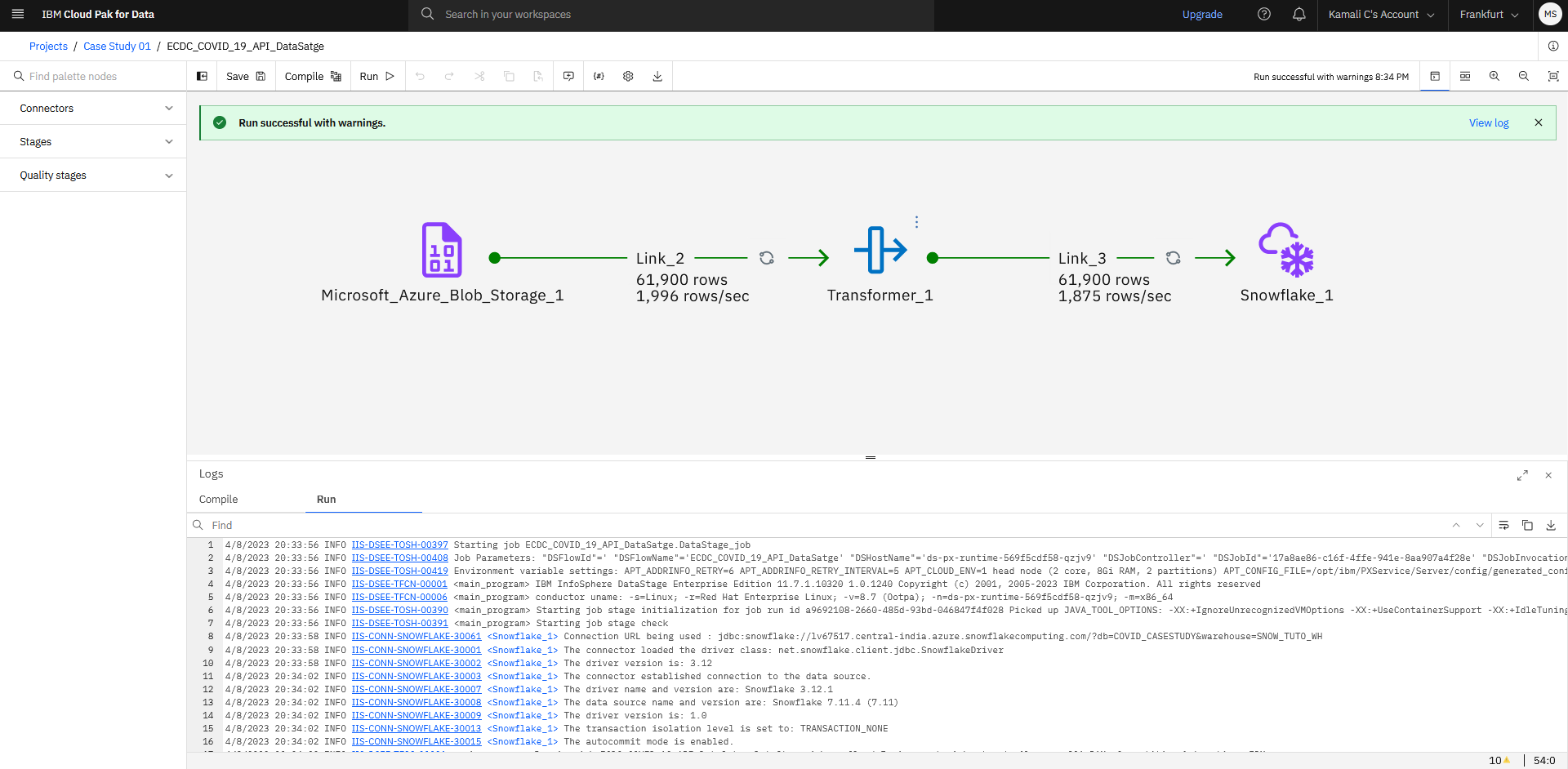
**Expression used in the Year\_Month column:**

“NumToStr(Link\_2.year, 0) : '\_' : NumToStr(Link\_2.month, 0)”

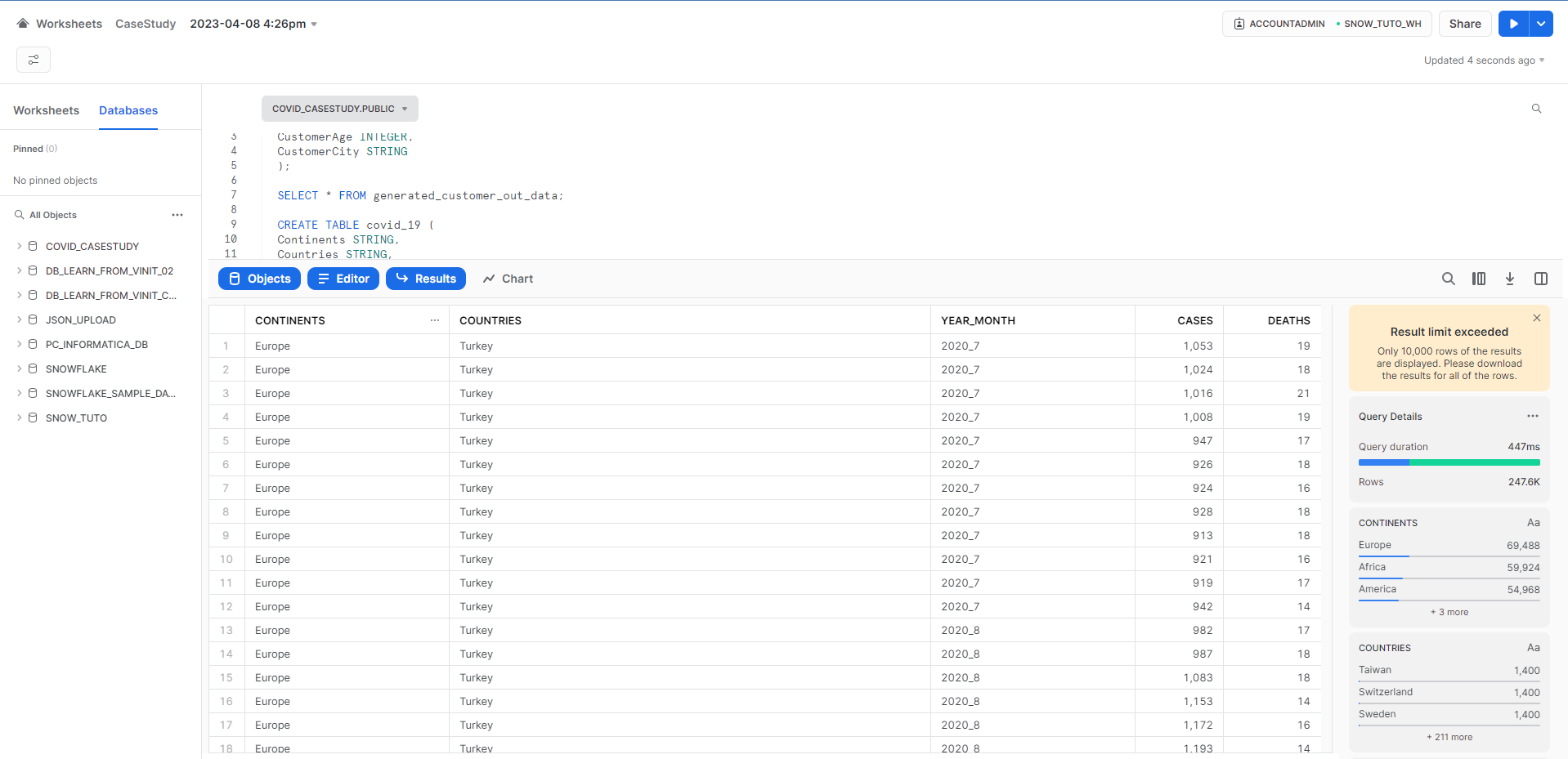
****

**Job Run:**

**Job run status. It resulted around 61,900 rows to snowflake.**

****

**Snowflake Table Preview [Loaded Data]:**

****