**Implementation Plan**

**Date:11/14/2024**

**Version: 1.0**

**Developer: Laxmi Dhungana**

**Status:Passed- Good to proceed.**

**Work flow**

The file was given for

a.Patient

b.Claims

* The task is to load those files in S3 bucket in input folder.
* Then read those file in data bricks and clean
* Once data is clean load those clean data into redshift table or load into s3 in output folder.

**Implementation and snapshot**

**Patient:**

**Step 1:**

**Created one folder for capstone project under s3 and subfolder named: Input. Loaded all the file inside the input folder.**

**A screenshot of a computer

Description automatically generated**

**Step 2:**

**Hadoop configuration is set in databricks**

**A screen shot of a computer

Description automatically generated**

**Step 2:**

**Read the data from s3 filepath from databricks.**

**A screenshot of a computer

Description automatically generated**

**Step 3: Checking duplicate**

**Checked duplicate running for some unique columsn. And if duplicate exist need to drop.**

**Since there was no an duplicate I didn’t run drop duplicate syntax**

**A screenshot of a computer program

Description automatically generated**

**Step 4:**

1. **Cleaning**

**Checking null if exist, if it does exist update only those columns NULL value to “NA”**

**A screenshot of a computer

Description automatically generated**

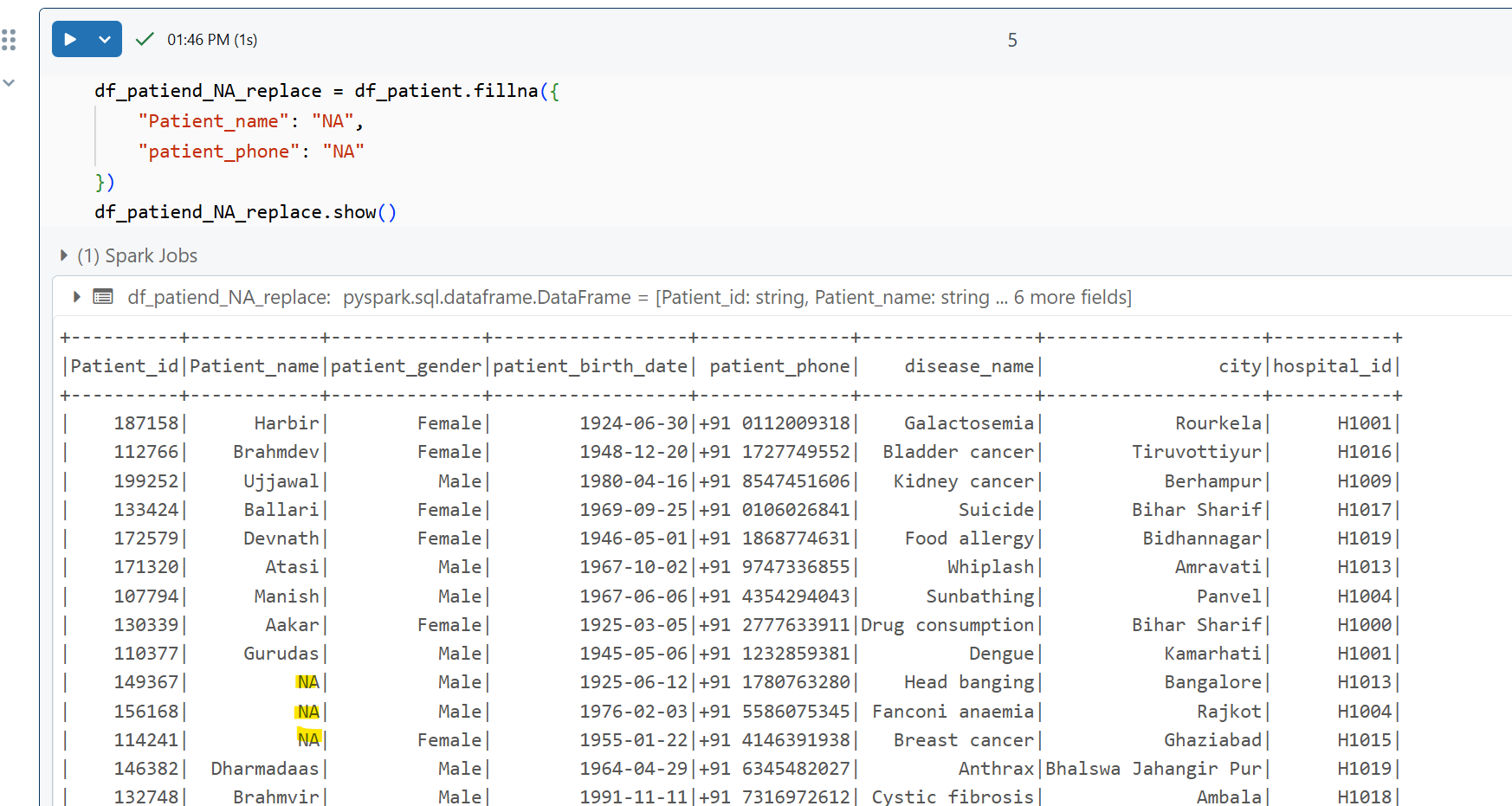
**Isnull check if there is null in columns.**

**Casting is used to convert Booleans to int as 1/0.**

**Sum is to count the number of null values in each columns**

1. **Cleaning**

**Updating those null value with “NA”**

****

**Fillinna() – this is used to fill null value with specific value given. For our requirement its “NA”**

**Step 5: This is the main requirement for using transformation:**

**Requirement: List of all the patients below age of 18 who admit for cancer.**

Our final detest after deleting dubs and change null value to NA is “df\_patiend\_NA\_replace”

Using this data frame, I filtered only those disease names which have cancer.

And for calculating the age of the patient.

I first calculated the numbers of month between current date and date of birth.

Then I divided the output I got by 12 , which can convert it to years.

Floor will get the full year by rounding down.

Then selected the age as needed.

Since there was no data for patients below 18 , I just have age 70 to test if code was working as expected.

**Below 18**

A screenshot of a computer

Description automatically generated

**Below 70**

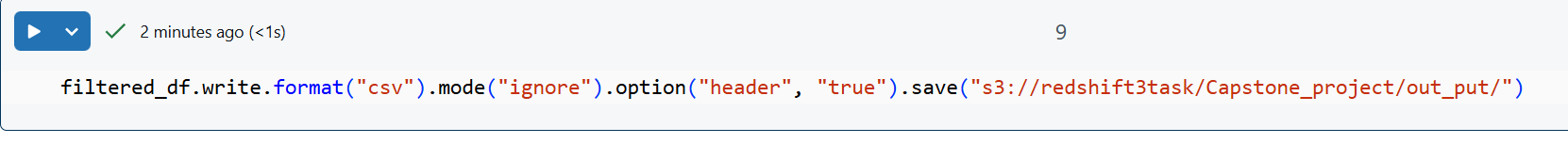
**A screenshot of a computer

Description automatically generated**

**Step 6 : Final data load.**

**We have 2 option to load cleaned data one in redshift and other in output folder in s3.**

**Since my Redshift threw some connection issue , I successfully able to load final data in s3 folder.**

****

**A screenshot of a computer

Description automatically generated**

**Step 7: Testing before pushing to QA.**

I downloaded the final output file in excel and checked if it all meet the requirements.

Checked if disease name have any other then cancer.

Checked and calculated the age and make sure

A screenshot of a computer

Description automatically generated