**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”**

**There was no null in any columns in claims data.**

**A screenshot of a computer code

Description automatically generated**

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

1. **Cleaning**

**Since there is No null value I didn’t run any code.**

**If there was null there the code looks like:**

**A screen shot of a computer code

Description automatically generated**

**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 patients who have cashless insurance and have total charges greater than or equal for Rs. 50,000.**

Here we need to pull patients’ names and details from patients table and the claims amount from claims data.

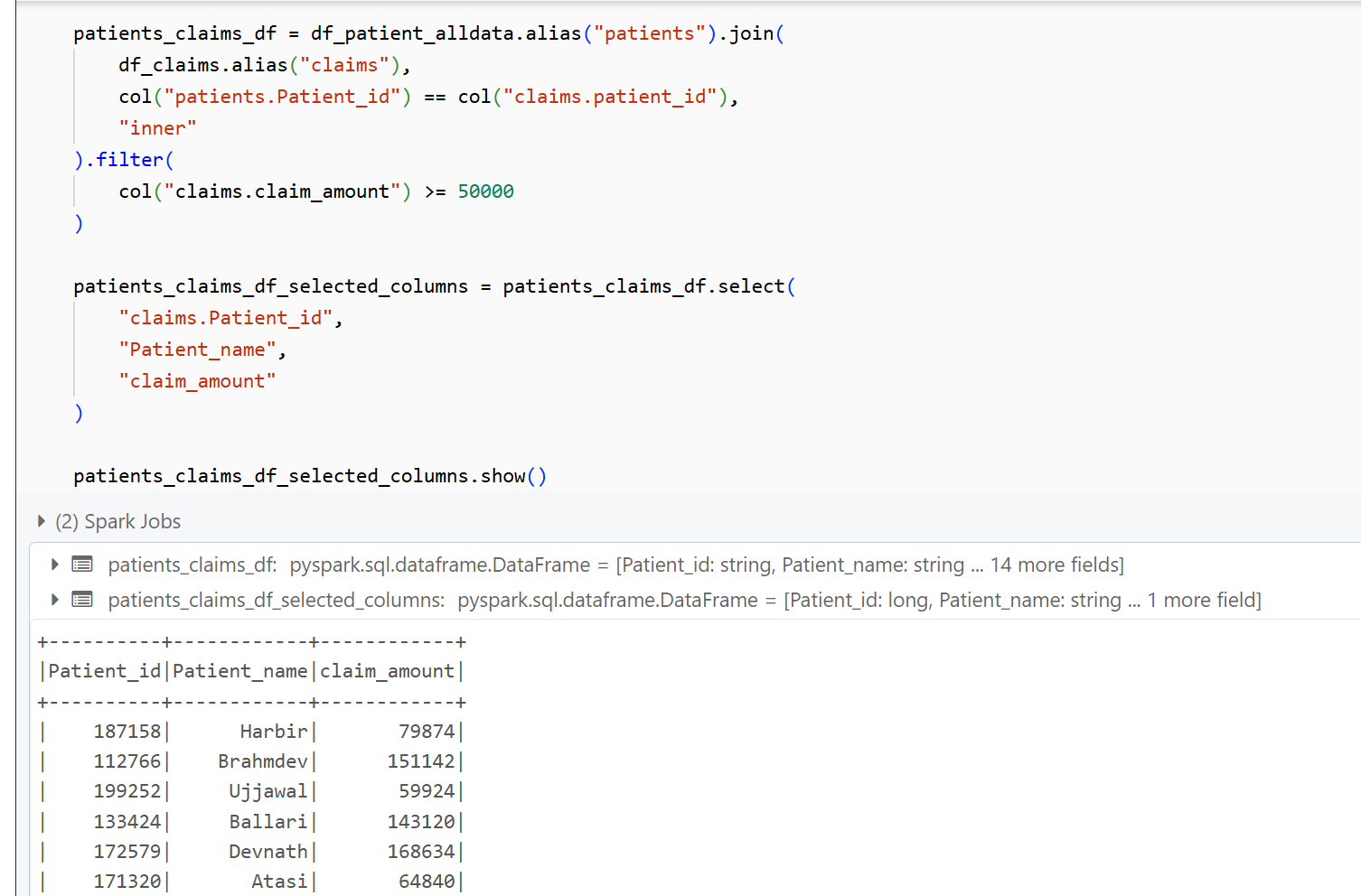
So, we need to join 2 tables on patients

Workflow:

First join 2 dataframe : claims and patients on patients\_id.

Then filter to pull the claims amount greater then re.50000.

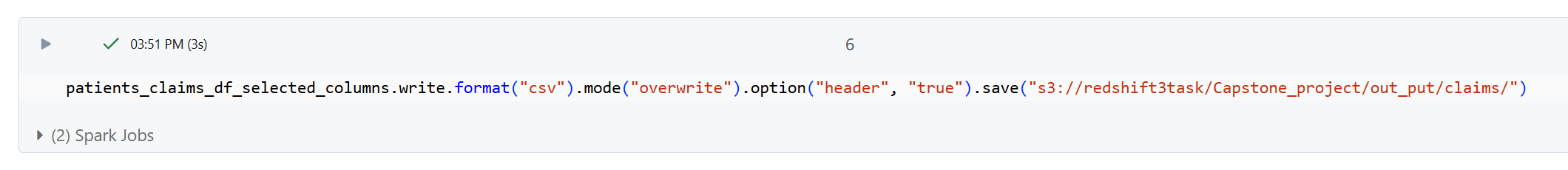
Then using select function selected only required column. In my case I just selected patients\_id, patients\_name and claim\_amount.



**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**

**File loaded sucessfully**

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

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

Make sure there is no null value

Check claims amount to make sure all data have greater then 50k

A screenshot of a computer

Description automatically generated

A screenshot of a spreadsheet

Description automatically generated