**SNOWFLAKE PROJECT DOCUMENTATION**

* **CAMP Batch B1 V3 – I**
* **Group 2**
* **Supervised by Mr. Pritam**

**Team Members: -**

Yashaswi Singh

Mohit Agarwal

Ritika Das

Stuti Bansal

Anivesh Gupta

Varan Kumar Gupta

Devashish Singh

**INDEX**

1. Project Description
2. Unit Test Report

* Creation of database
* Creation of schemas
* Creation of table as per dataset
* Creation of Integration Object
* Creation of external stage for loading the data structure
* Creation of snowpipe for autoingesting of data from S3 bucket
* Creation of stream on the given table
* SCD 2 operation on the consumer table

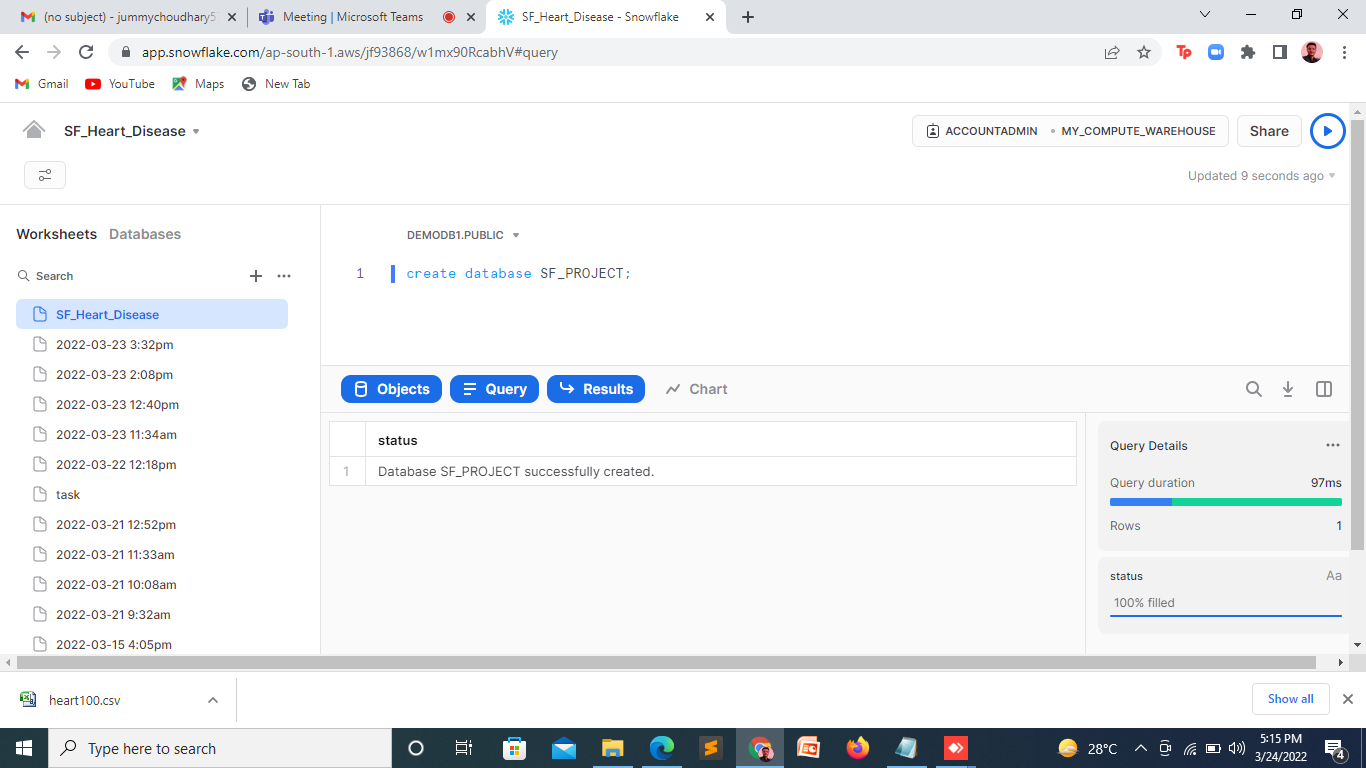
1. Data Analysis

**Project Description:** This project is to ingest and analyze a [dataset](https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease) from kaggle having details related to causes which can lead to heart disease.

**Unit Test Report: -**

**Query:** Create database SF\_PROJECT;

**Description:** Created a database name SF\_PROJECT



*Fig 1*

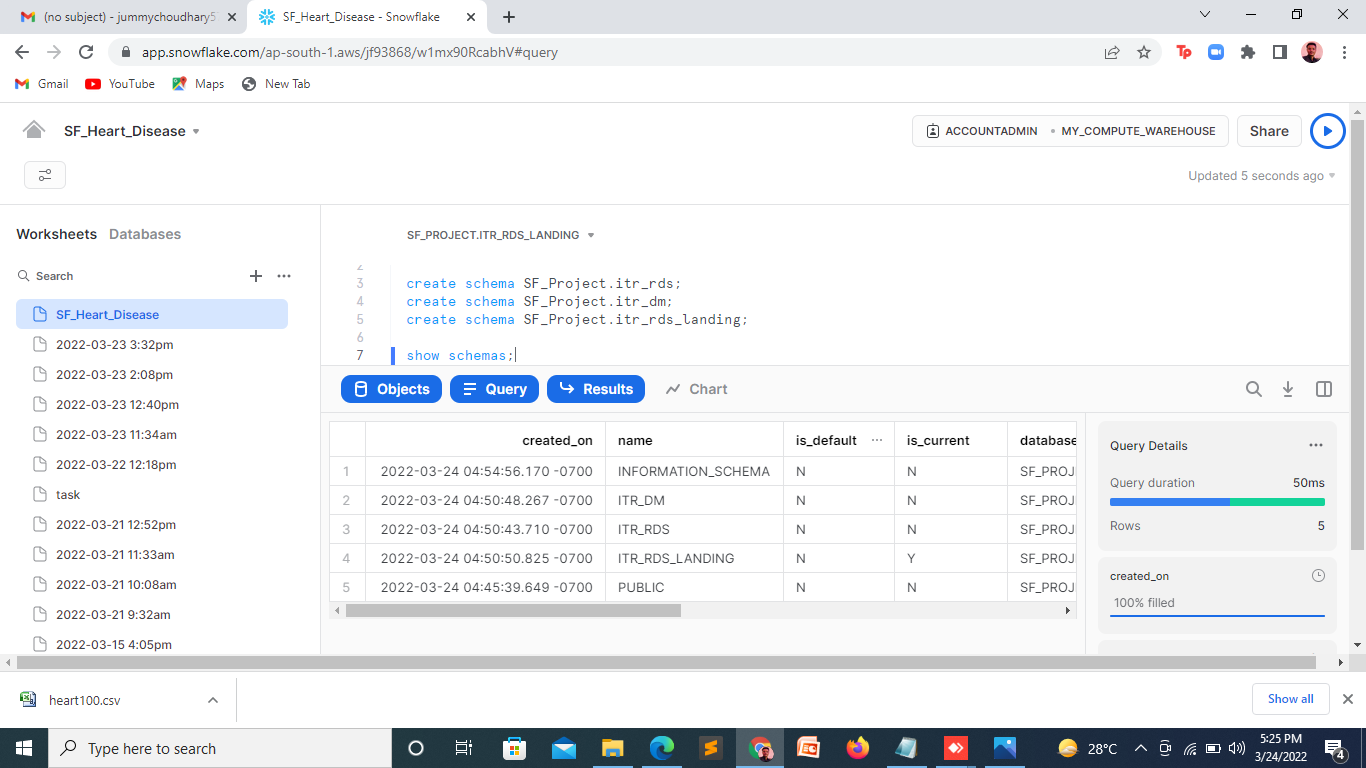
**Query:** create schema SF\_Project.itr\_rds;

create schema SF\_Project.itr\_dm;

create schema SF\_Project.itr\_rds\_landing;

**Description:** Created 3 schemas namely ITR\_RDS, ITR\_DM, ITE\_RDS\_LANDING

*\*SCHEMA: A database schema defines how data is organized within a relational database.*

**

*Fig 2*

**Query:** create table heart\_disease (

heartdisease varchar,

BMI decimal,

smoking varchar,

AlcoholDrinking varchar,

Stroke varchar,

PhysicalHealth number,

MentalHealth number,

DiffWalking varchar,

Sex varchar,

AgeCategory varchar,

Race varchar,

Diabetic varchar,

PhysicalActivity varchar,

GenHealth varchar,

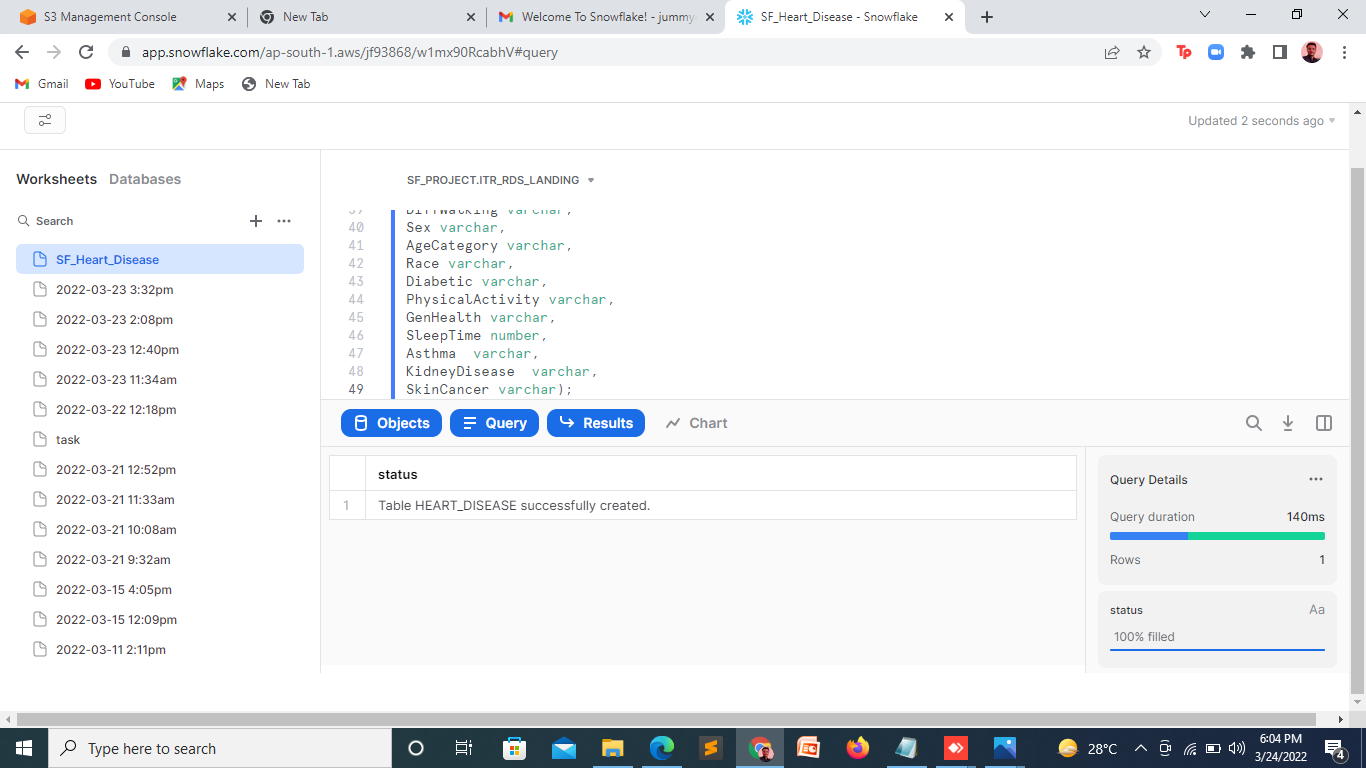
SleepTime number,

Asthma varchar,

KidneyDisease varchar,

SkinCancer varchar);

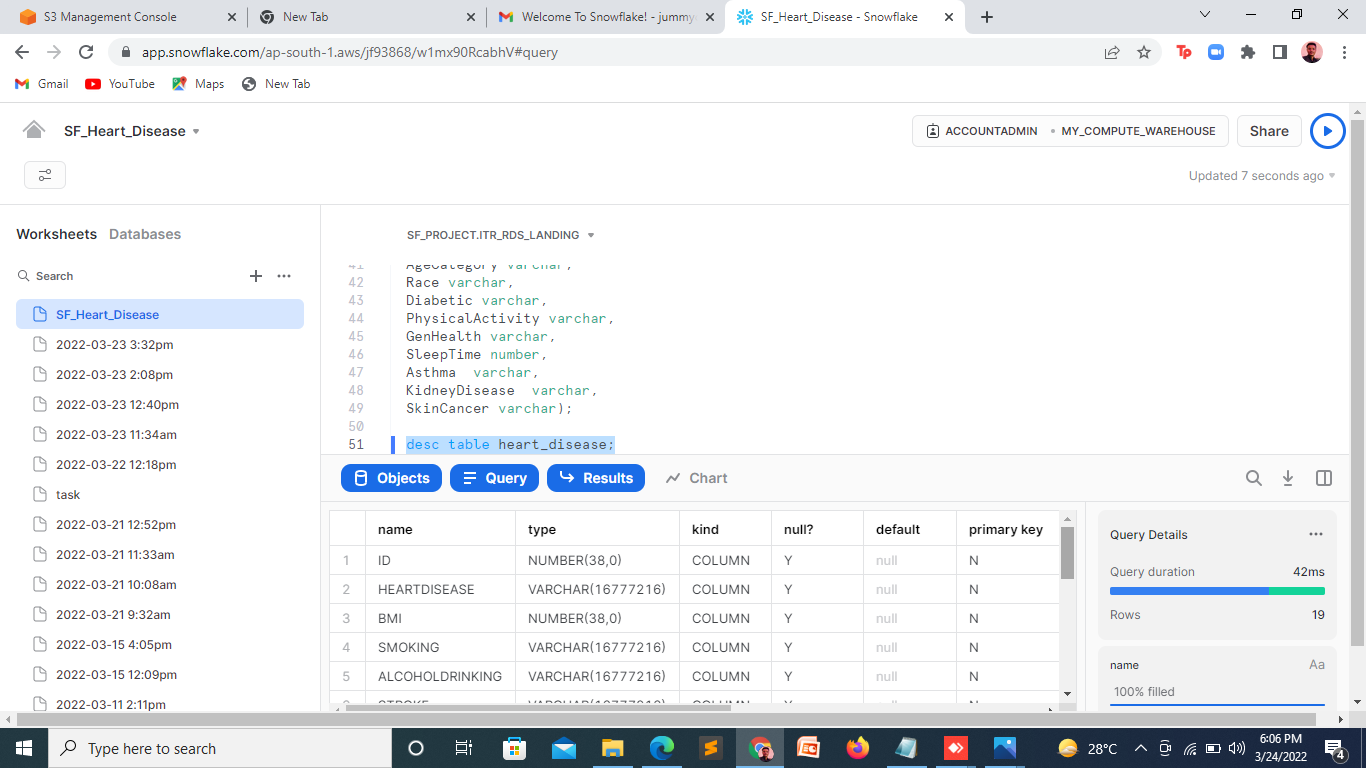
**Description:** Table is created with the desired columns and its specified datatype.



*Fig 3*

**Query:** desc table heart\_disease;

**Description:** Showing the details of the table



*Fig 4*

**Query:** create or replace storage integration s3\_int\_obj

type = external\_stage

storage\_provider = s3

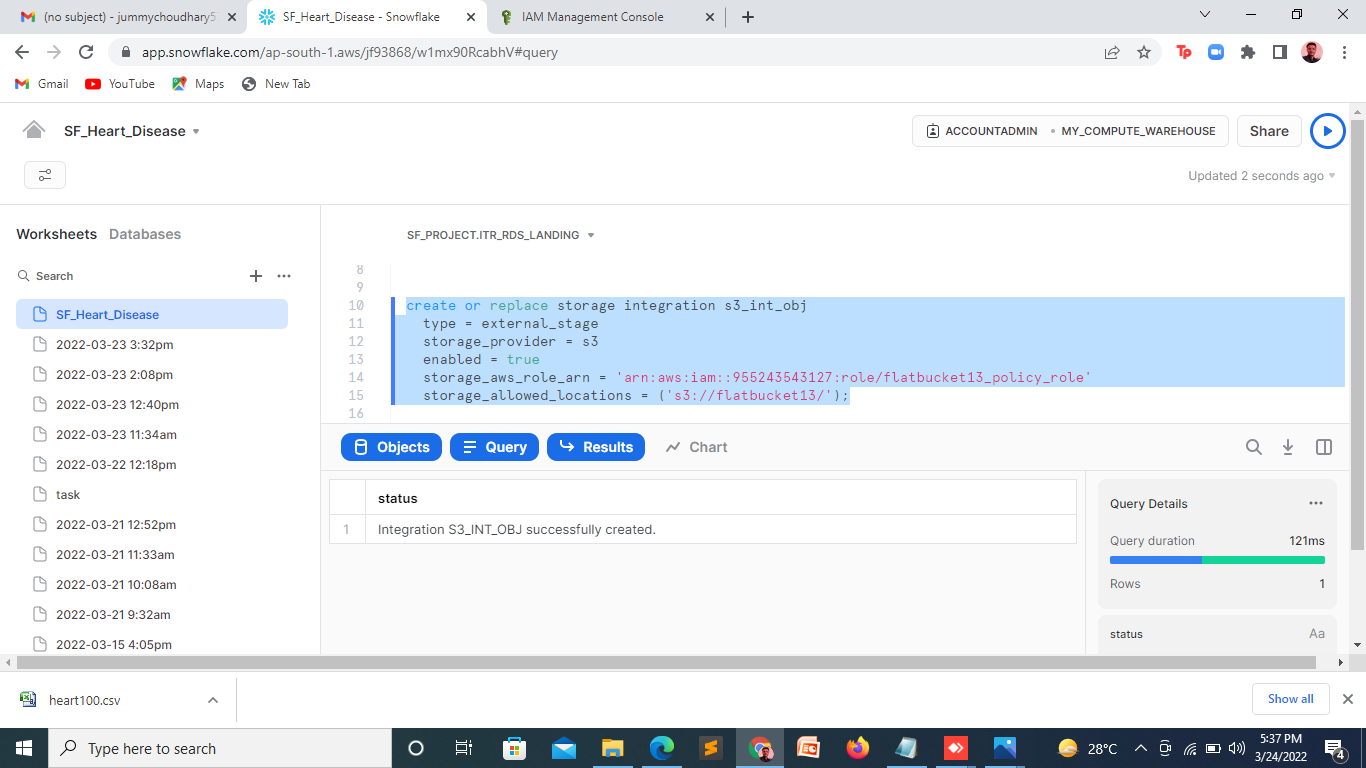
enabled = true

storage\_aws\_role\_arn = 'arn:aws:iam::084370864130:role/flatbucket6\_policy\_role'

storage\_allowed\_locations = ('s3://flatbucket6/');

**Description:** Connecting snowflake with our s3 bucket named FLATBUCKET6 in AWS.

*\*Storage integration: A storage integration is a Snowflake object that stores a generated identity and access management (IAM) entity for your external cloud storage, along with an optional set of allowed or blocked storage locations (Amazon S3, Google Cloud Storage, or Microsoft Azure).*



*Fig 5*

**Query:** create or replace stage sf\_project.itr\_rds\_landing.my\_ext\_stage

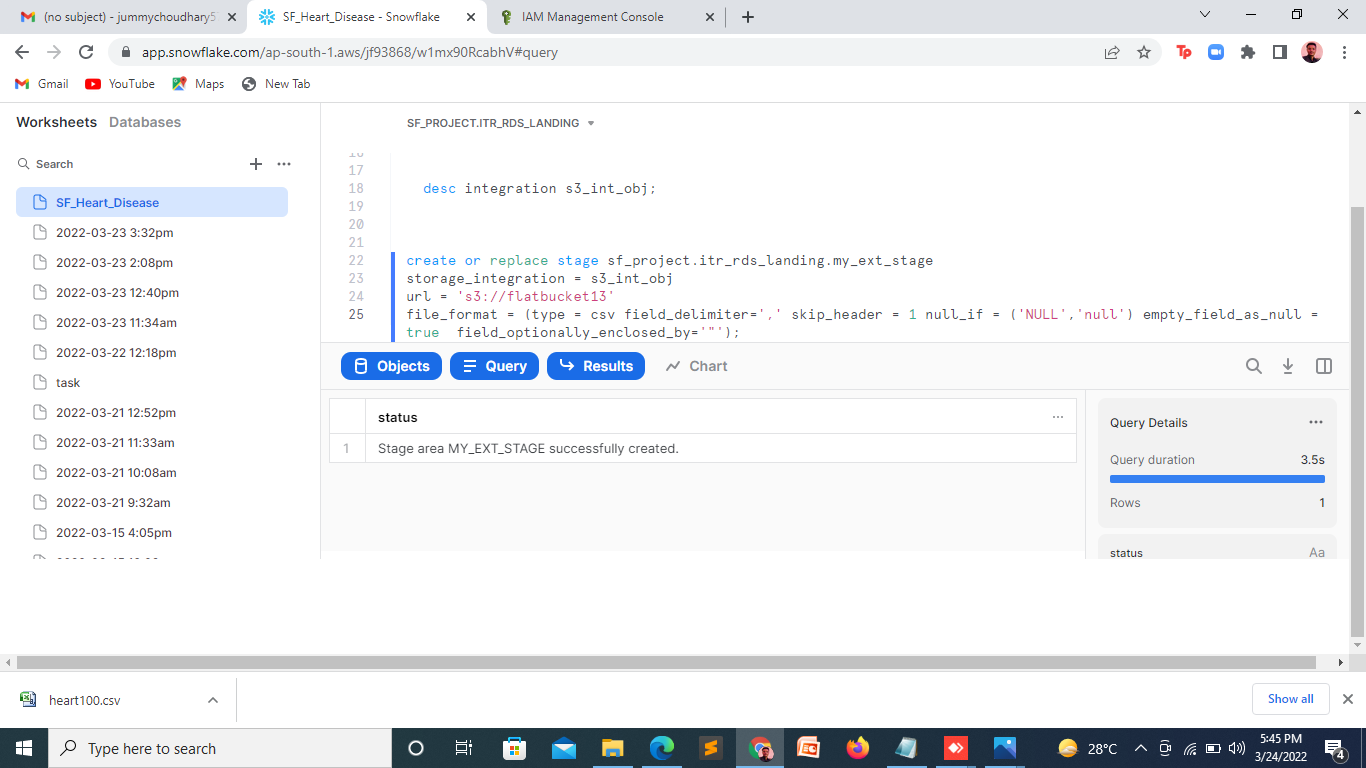
storage\_integration = s3\_int\_obj

url = 's3://flatbucket12'

file\_format = (type = csv field\_delimiter=',' skip\_header = 1 null\_if = ('NULL','null') empty\_field\_as\_null = true field\_optionally\_enclosed\_by='"');

**Description:** Here an external stage is created named as MY\_EXT\_STAGE with storage integration as S3\_INT\_OBJ. The file is formatted in csv form with delimiter as “,” and with the help of skip\_header attribute as 1, the first row i.e. the header of the file will be skipped.

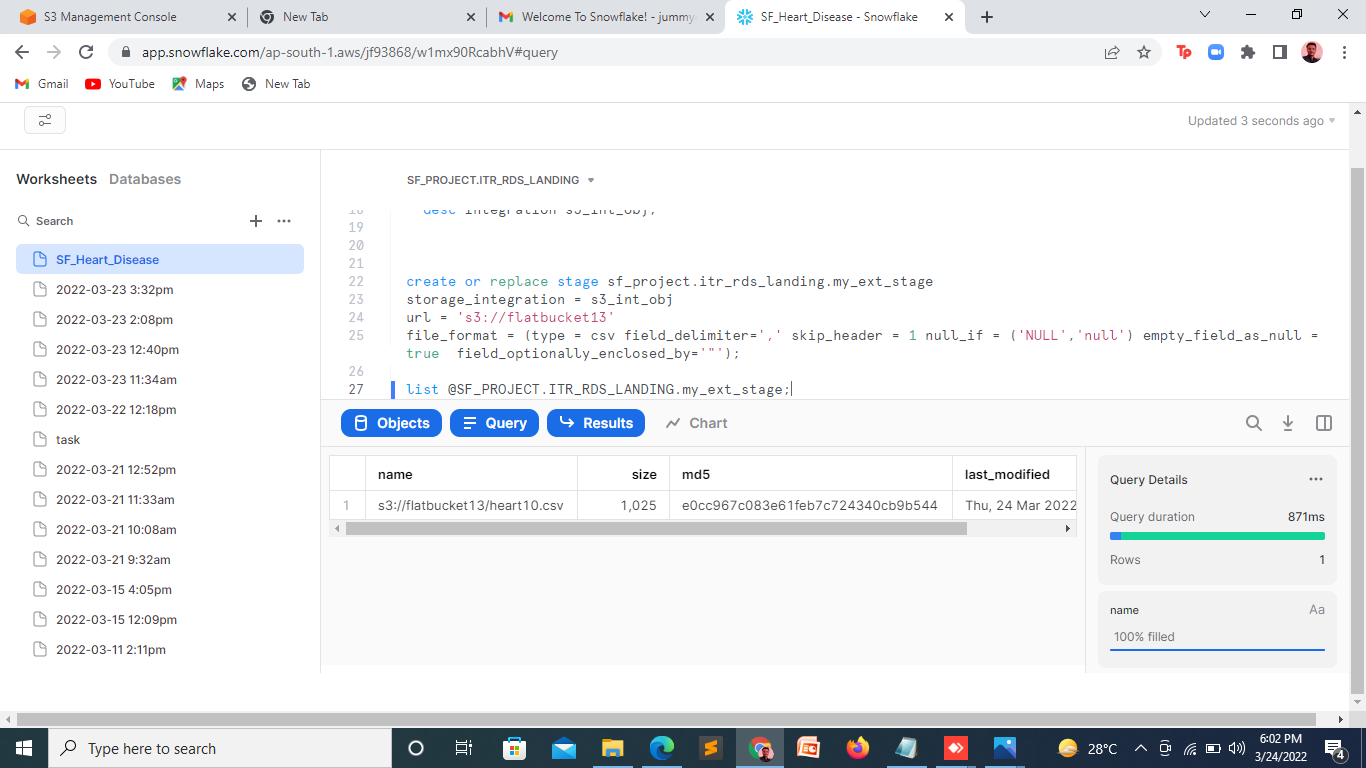
*\*External stage: Creates an interface between Snowflake and an external cloud storage location.*



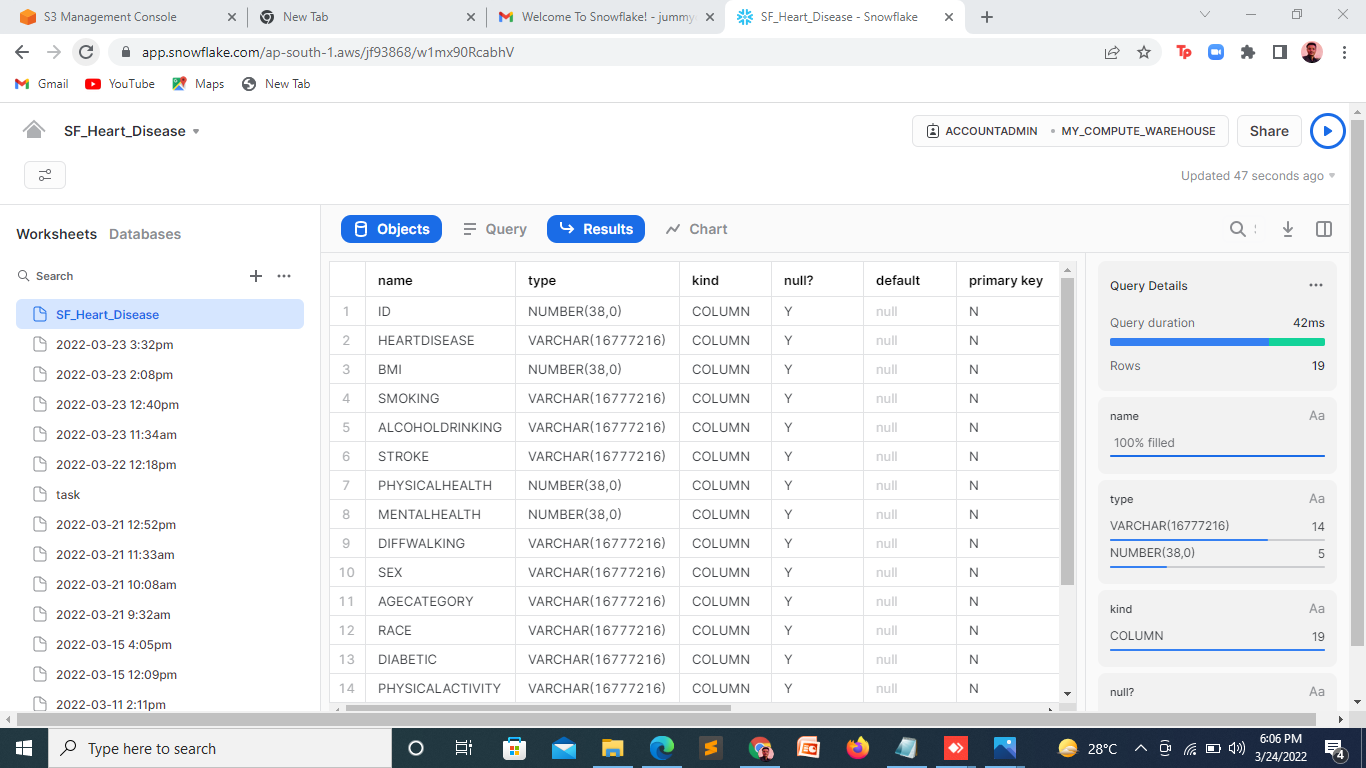
*Fig 6*

**Query:** list @sf\_project.itr\_rds\_landing.my\_ext\_stage;

**Description:** Listing the details of external stage.



*Fig 7.1*



*Fig 7.2*

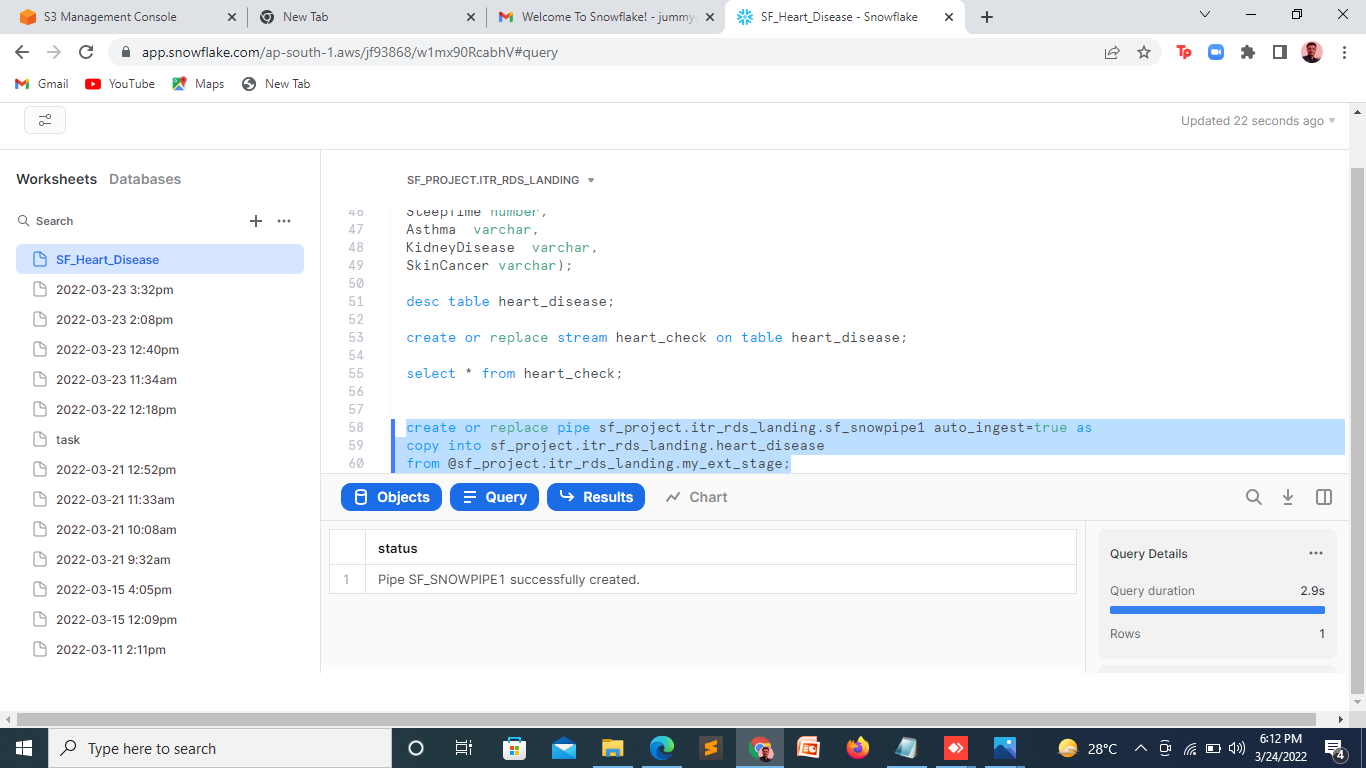
**Query:** create or replace pipe sf\_project.itr\_rds\_landing.sf\_snowpipe1

auto\_ingest=true as

copy into sf\_project.itr\_rds\_landing.heart\_disease

rom @sf\_project.itr\_rds\_landing.my\_ext\_stage;

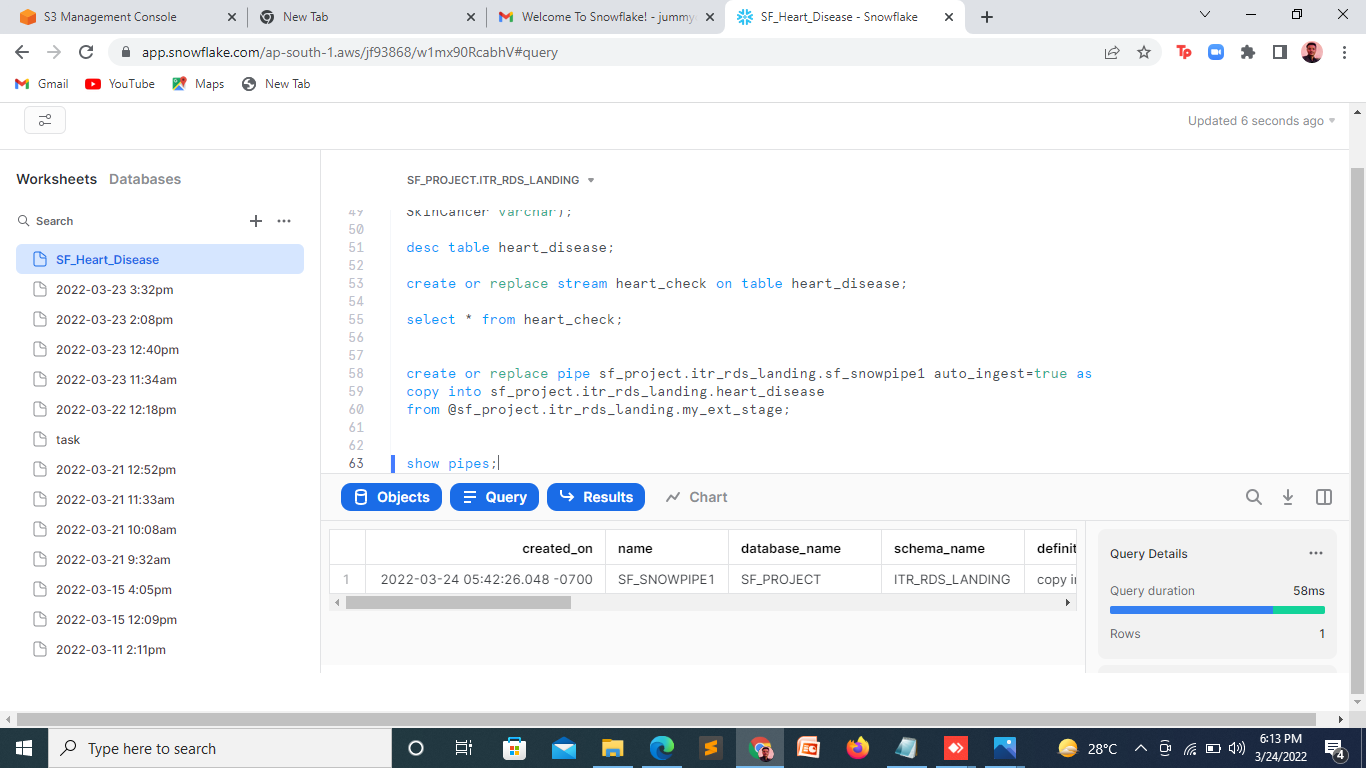
**Description:** Here a pipe is created named as SF\_SNOWPIPE1 given auto\_ingest as true i.e whenever new data is found in the MY\_EXT\_STAGE, then it is automatically inserted in the table.



*Fig 8*

**Query:** show pipes;

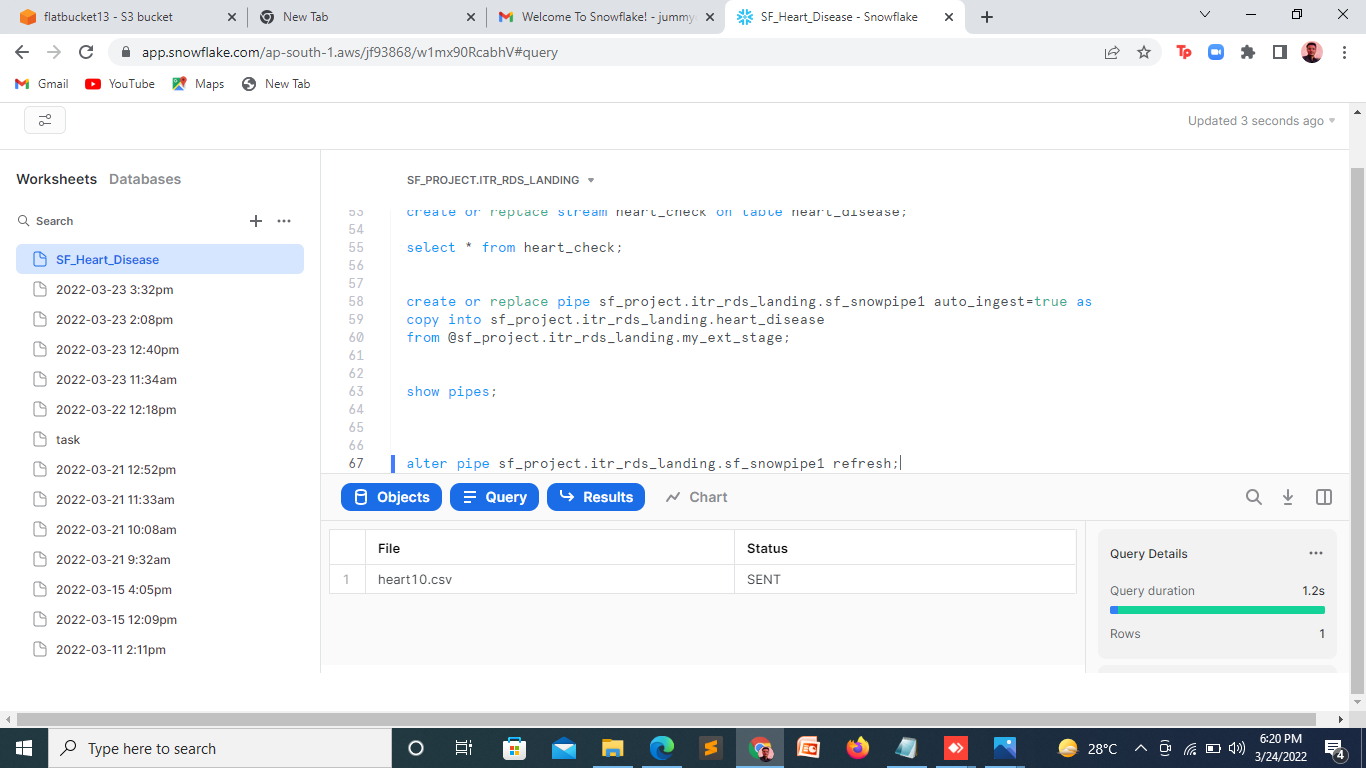
**Description:** Showing the description of pipes created above.



*Fig. 9*

**Query:** alter pipe sf\_project.itr\_rds\_landing.sf\_snowpipe1 refresh;

**Description:** Tells whether the dataset is sent or not.



*Fig. 10*

**Query:** create table heart\_disease\_tgt (

id number,

heartdisease varchar,

BMI decimal,

smoking varchar,

AlcoholDrinking varchar,

Stroke varchar,

PhysicalHealth number,

MentalHealth number,

DiffWalking varchar,

Sex varchar,

AgeCategory varchar,

Race varchar,

Diabetic varchar,

PhysicalActivity varchar,

GenHealth varchar,

SleepTime number,

Asthma varchar,

KidneyDisease varchar,

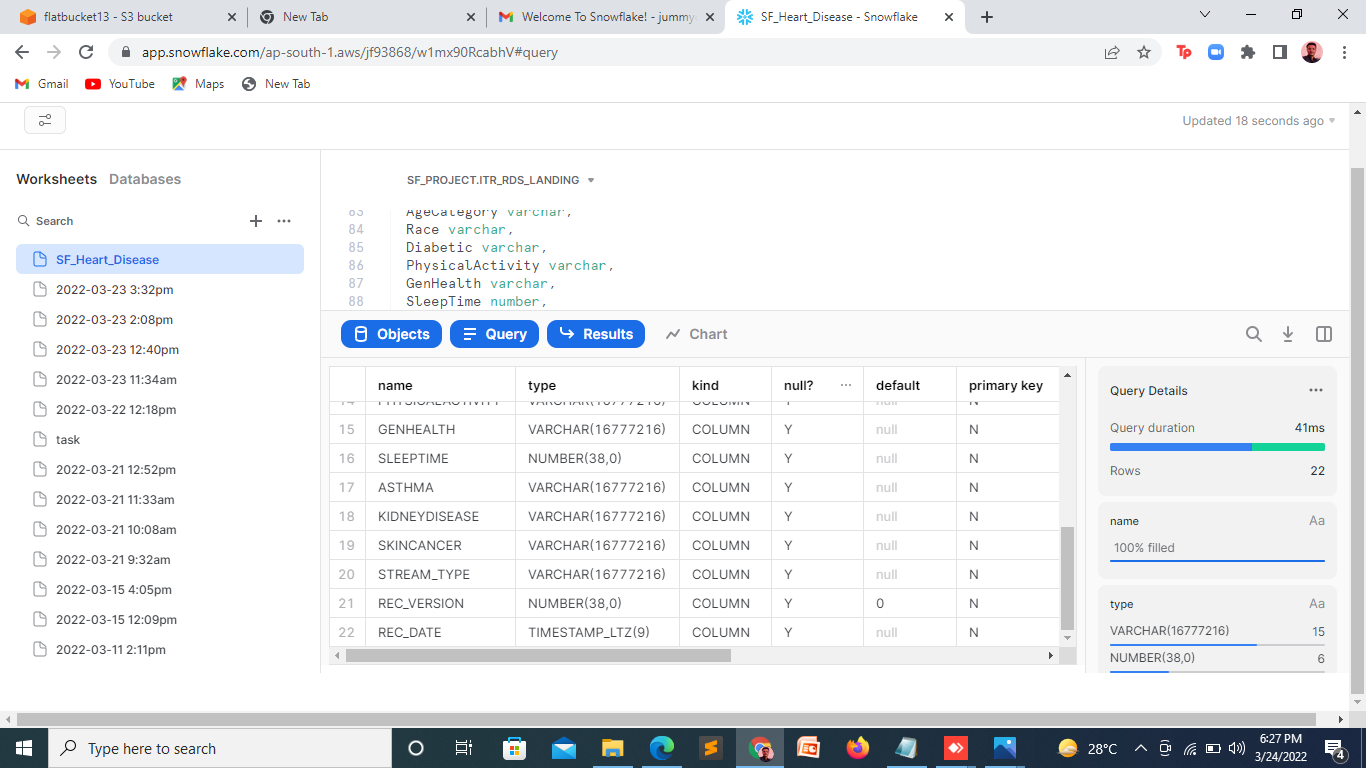
SkinCancer varchar,

stream\_type string default null,

rec\_version number default 0,

REC\_DATE TIMESTAMP\_LTZ);

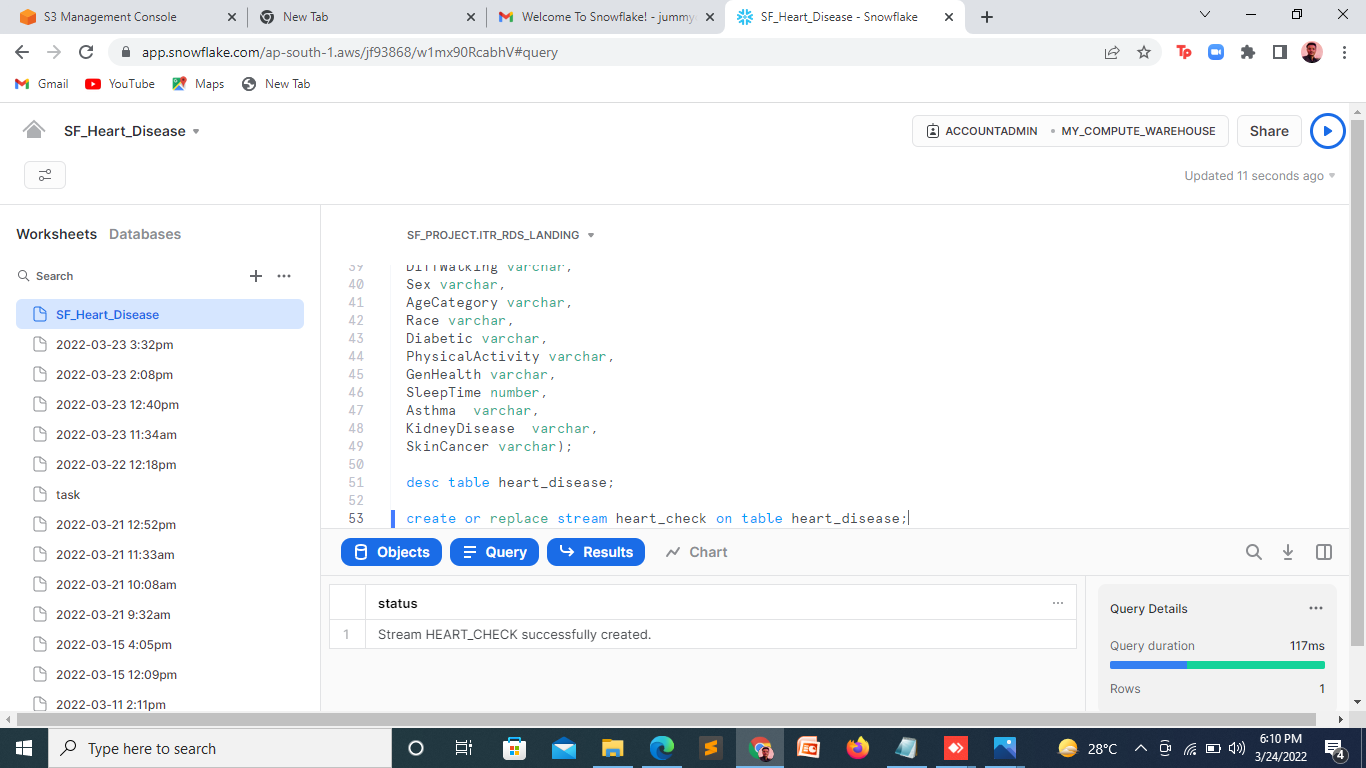
**Description:** Target table is created with the desired columns and its specified datatype.



*Fig. 11*

**Query:** create or replace stream HEART\_CHECK on table HEART\_DISEASE;

**Description:** Created a stream to update and load data from landing table to consumer table.



*Fig. 12*

**Query:** CREATE TASK heart\_task

WAREHOUSE = my\_first\_warehouse

SCHEDULE = '1 minute'

WHEN

SYSTEM$STREAM\_HAS\_DATA('heart\_check')

AS

merge into heart\_disease\_tgt t

using heart\_check s

on t.id=s.id and (metadata$action='DELETE')

when matched and metadata$isupdate='FALSE'

then update set rec\_version=9999,

stream\_type='DELETE' when matched

and metadata$isupdate='TRUE' then update set rec\_version=rec\_version-1

when not matched then

insert (id, heartdisease, BMI, smoking, Alcohol Drinking, Stroke, PhysicalHealth

, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic

, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer

, stream\_type, rec\_version, REC\_DATE)

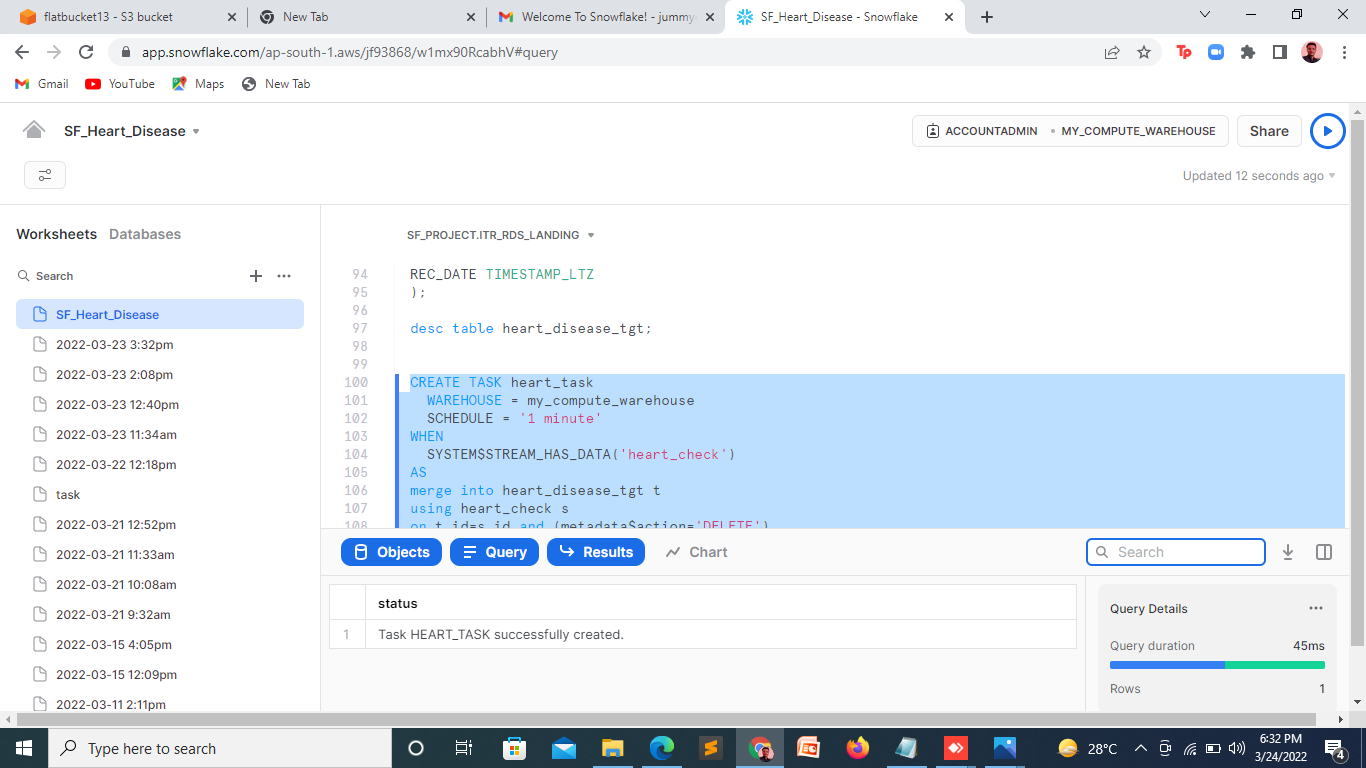
values(s.id, s.heartdisease , s.BMI ,smoking ,s.AlcoholDrinking ,s.Stroke

,s.PhysicalHealth ,s.MentalHealth ,s.DiffWalking ,s.Sex , s.AgeCategory ,s.Race

,s.Diabetic ,s.PhysicalActivity,s.GenHealth ,s.SleepTime,s.Asthma,s.KidneyDisease

,s.SkinCancer, metadata$action,0,CURRENT\_TIMESTAMP() );

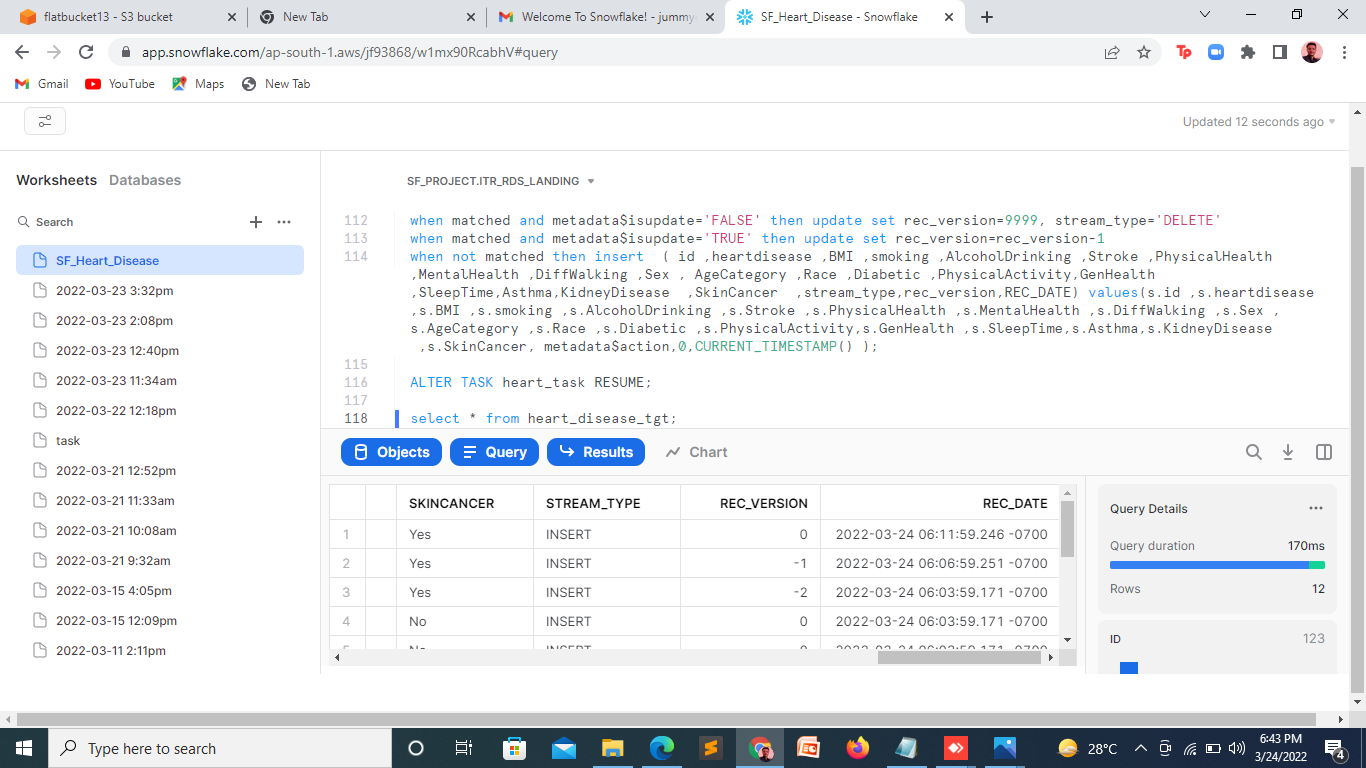
**Description:** A task is created to automate changes from landing table to consumer table and here we are using type 2 scd.



*Fig. 13*

**Query:** select \* from heart\_disease\_tgt;

**Description:** showing the whole target table i.e HEART\_DISEASE\_TGT at once.



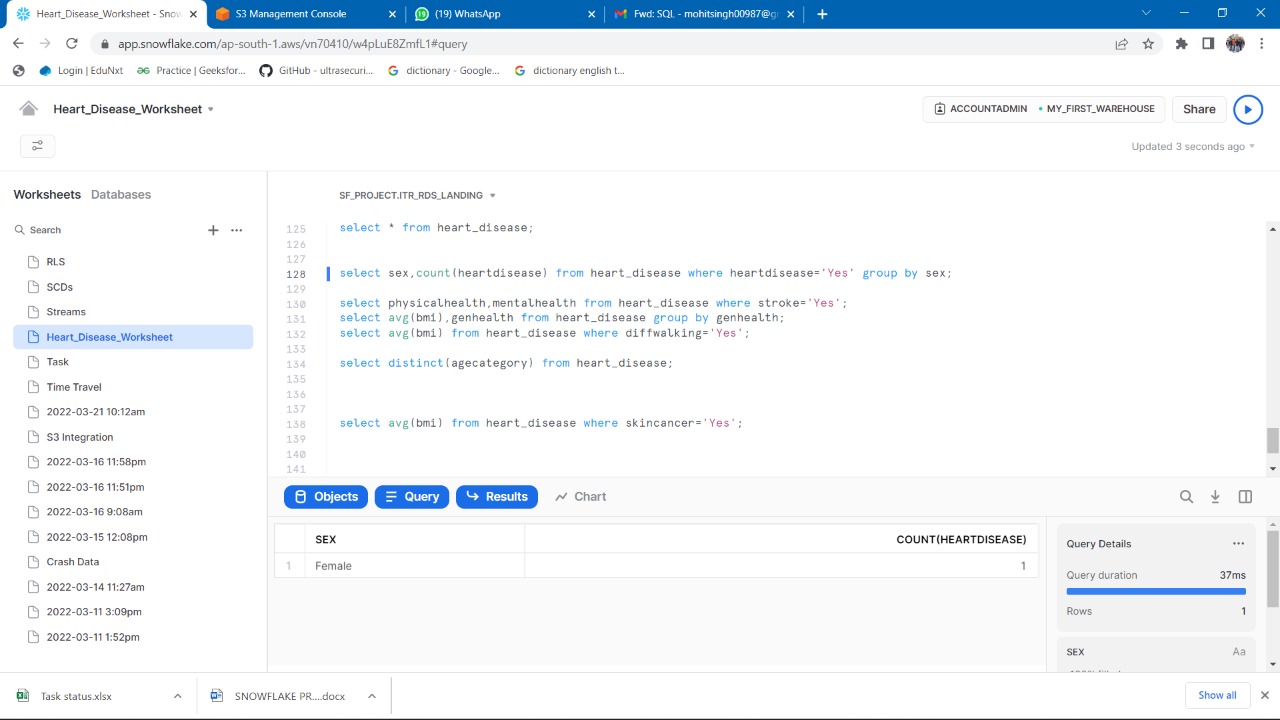
*Fig. 14*

**Data Analysis: -**

* Which gender have majority heart attack?

select sex, count(heartdisease) from heart\_disease where heartdisease='Yes' group

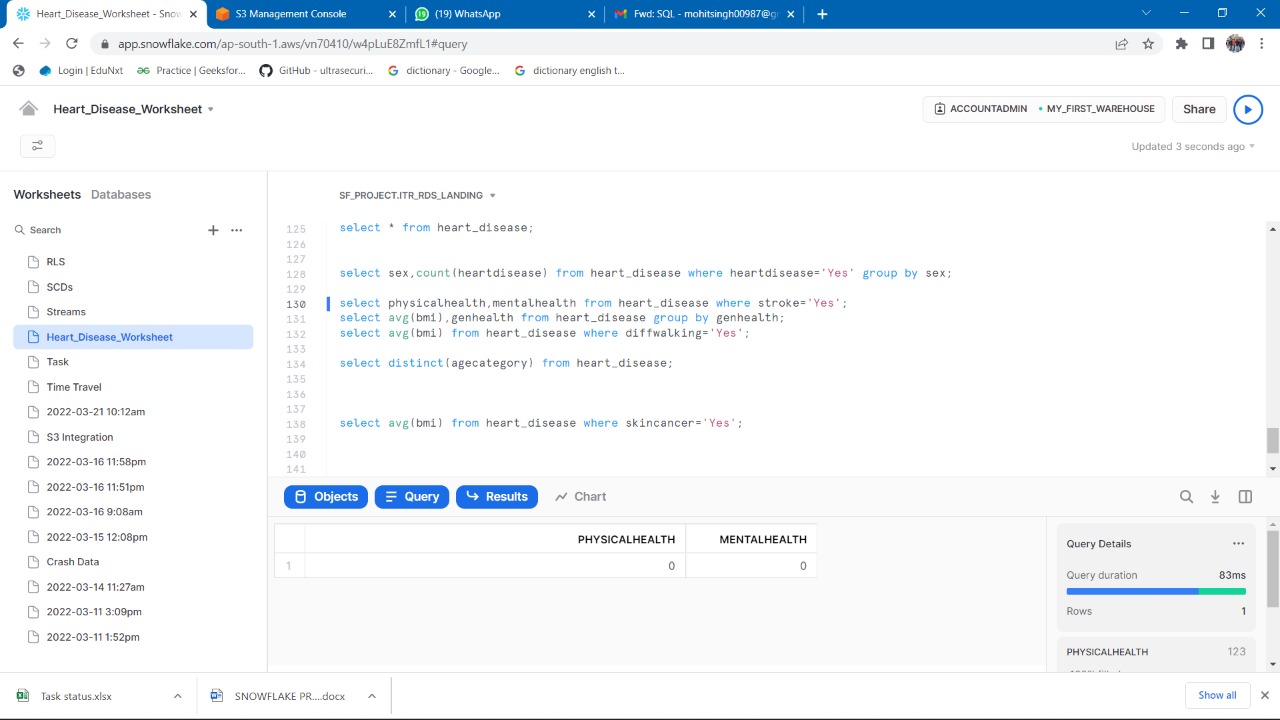
by sex;



*Solution 1*

* Which combination of physical and mental health causes stroke?

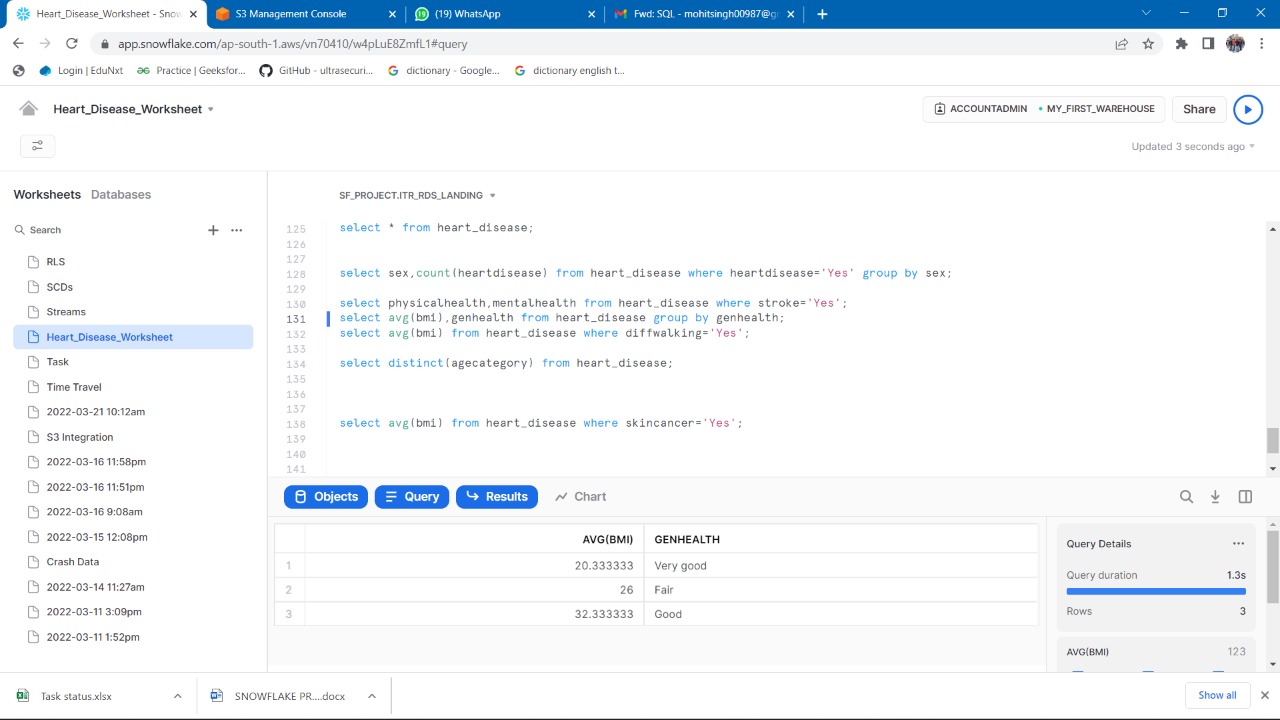
select physicalhealth, mentalhealth from heart\_disease where stroke='Yes';



*Solution 2*

* GenHealth wise average BMI of people?

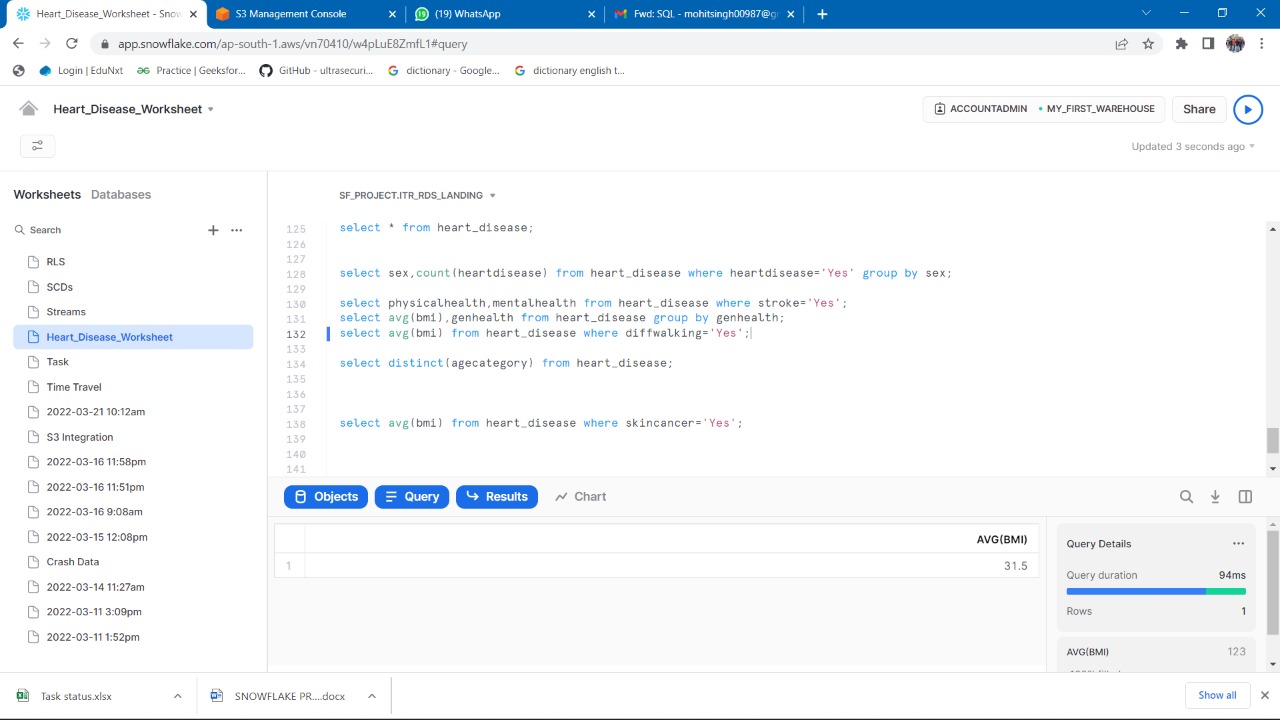
select avg(bmi),genhealth from heart\_disease group by genhealth;



*Solution 3*

* Average age of people having problem in walking?

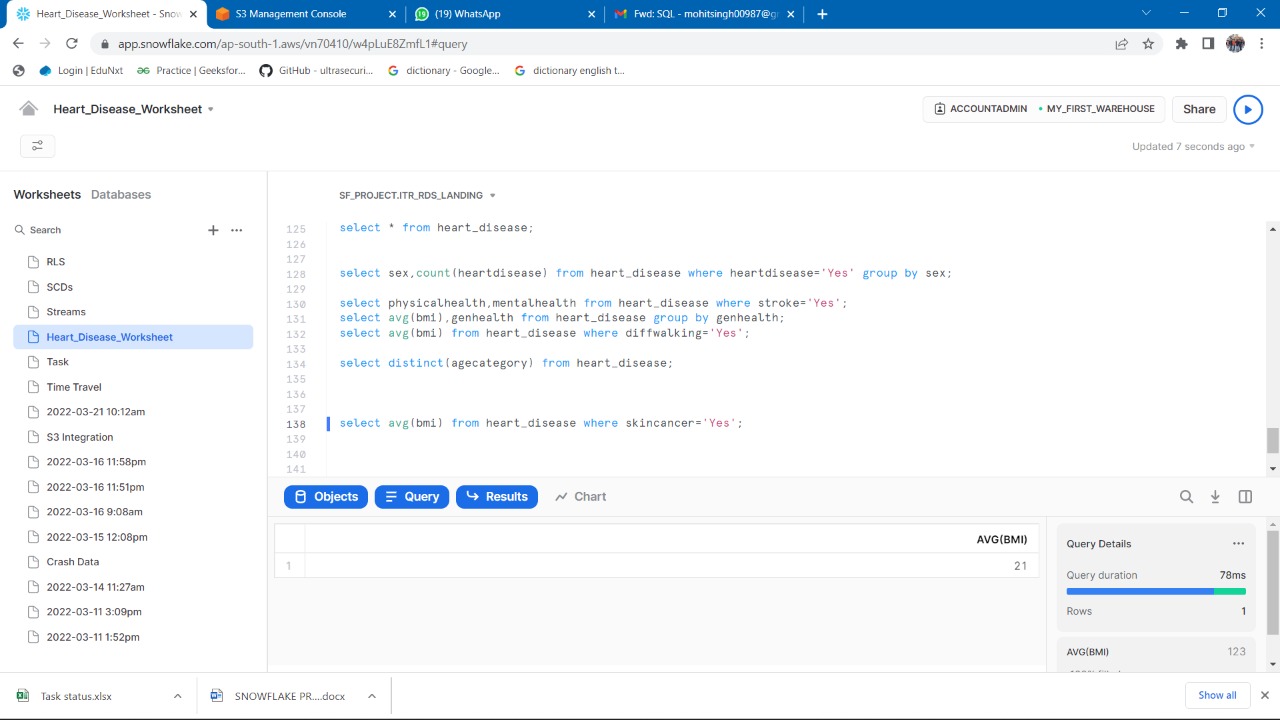
select avg(bmi) from heart\_disease where diffwalking='Yes';



*Solution 4*

* Average BMI of people having Skin Cancer

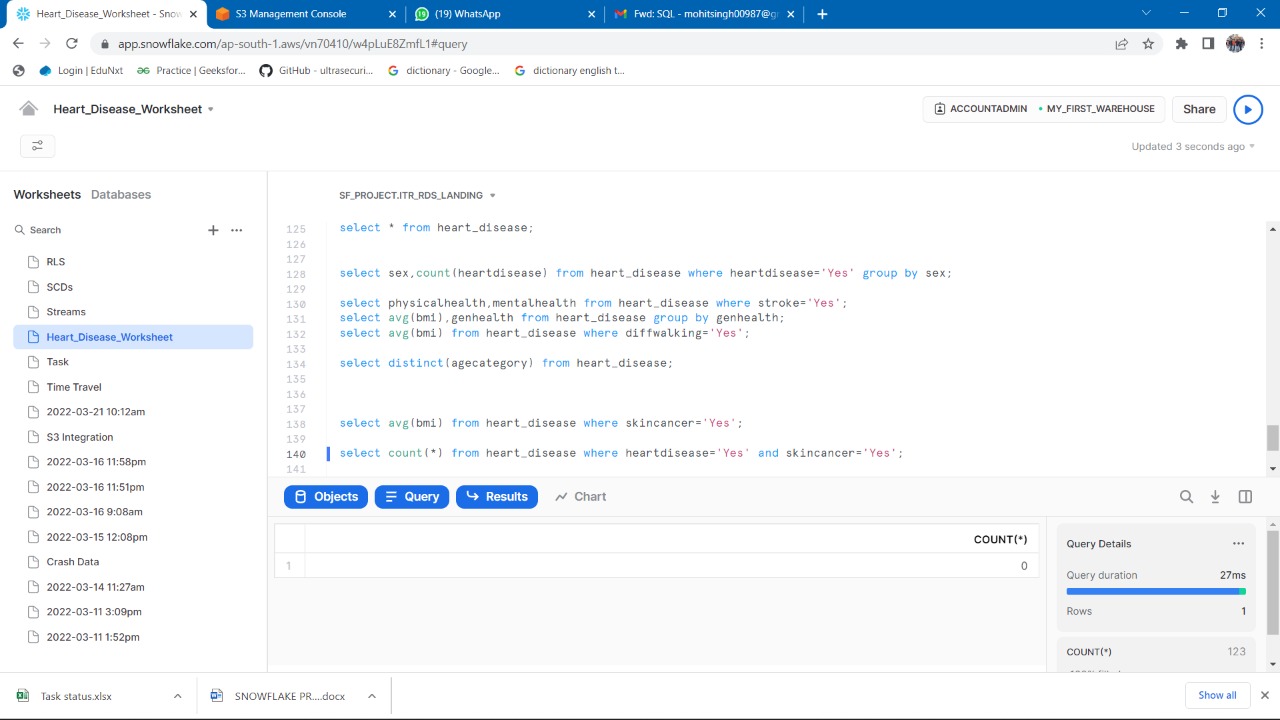
Select avg(bmi) from heart\_disease where skincancer='Yes';



*Solution 5*

* Possibilities/Percentage of people having both Heart Diseases stroke and Skin Cancer

Select count(\*) from heart\_disease where heartdisease='Yes' and skincancer='Yes';



*Solution 6*