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

Project Description: This project is to ingest and analyze a dataset from kaggle having details related to causes which can lead to heart disease.

Tasks Performed: -

Query: Create database SF_PROJECT;
Description: Created a database name SF_PROJECT

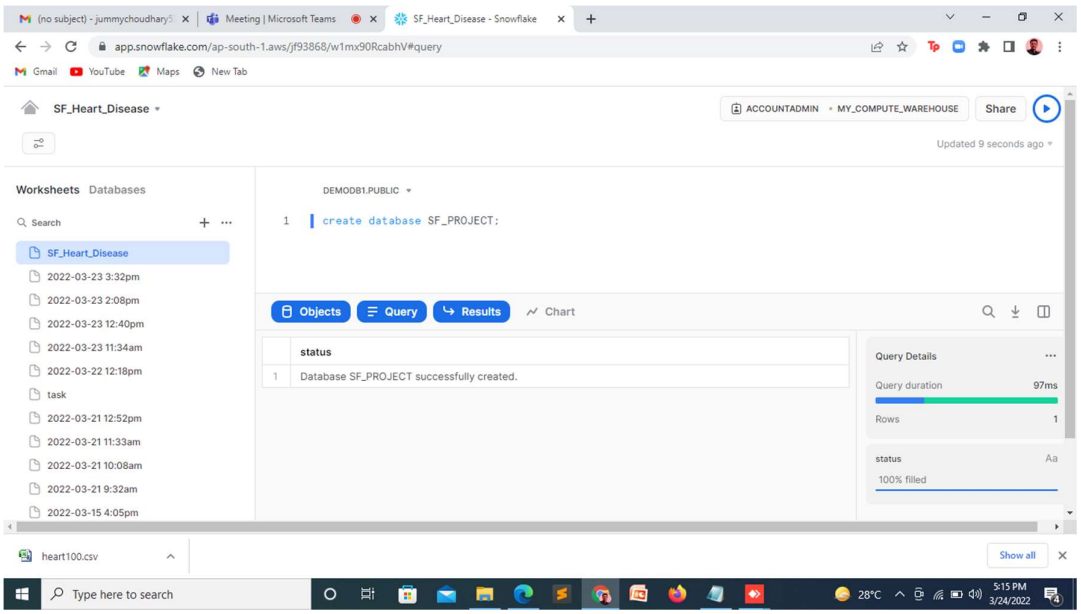


Fig 1

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.*

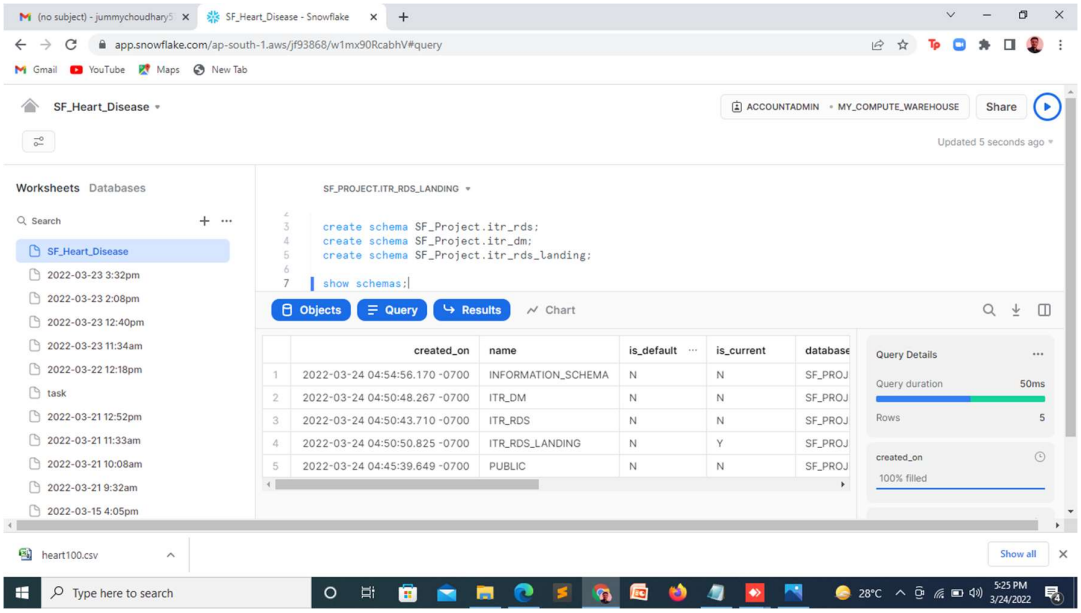


Fig 2

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.

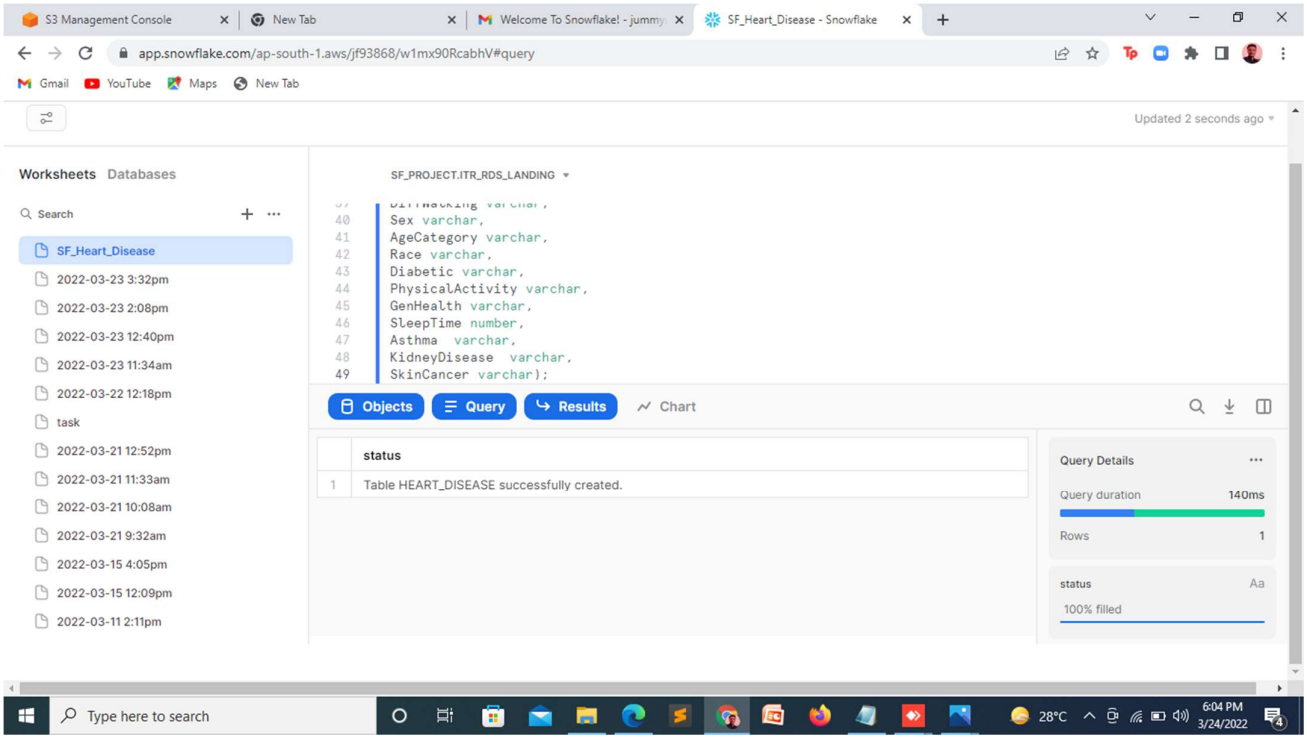


Fig 3

Query: desc table heart_disease;

Description: Showing the details of the table

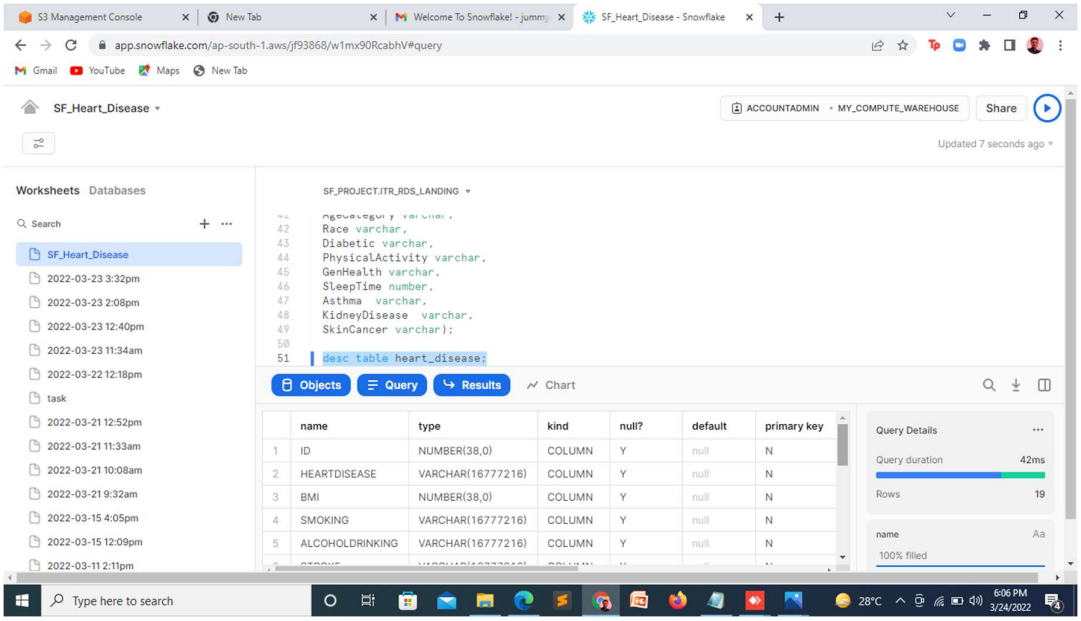


Fig 4

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).*

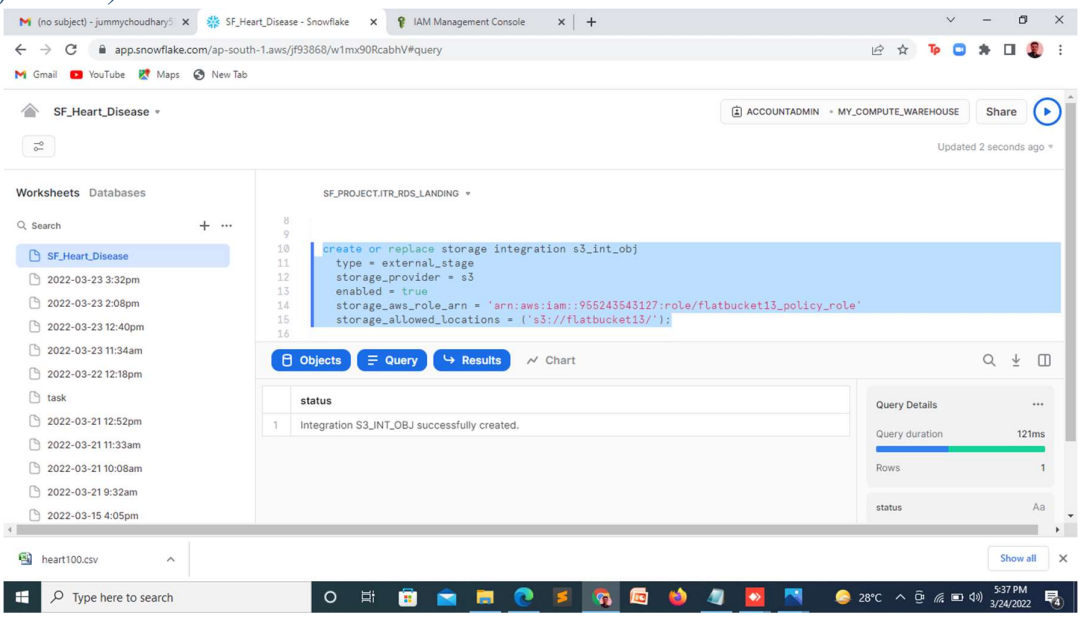


Fig 5

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.*

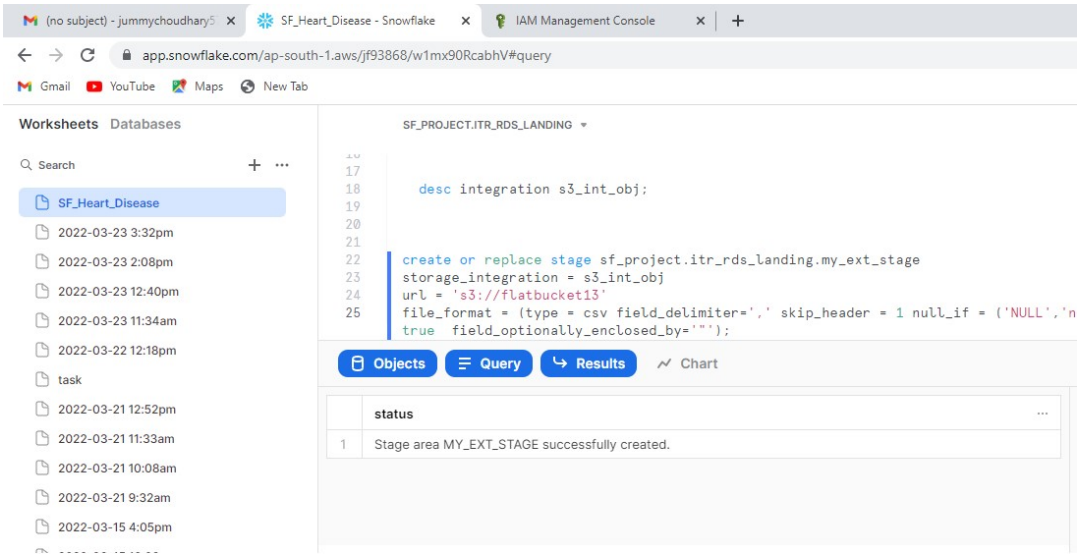


Fig 6

Query: list @sf_project.itr_rds_landing.my_ext_stage;
Description: Listing the details of external stage.

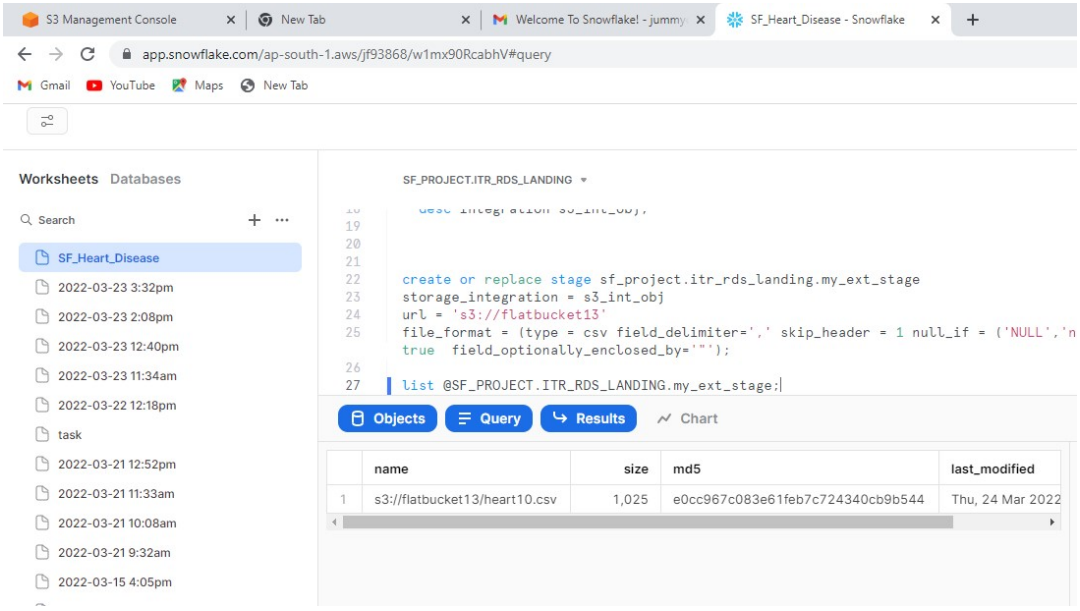


Fig 7.1

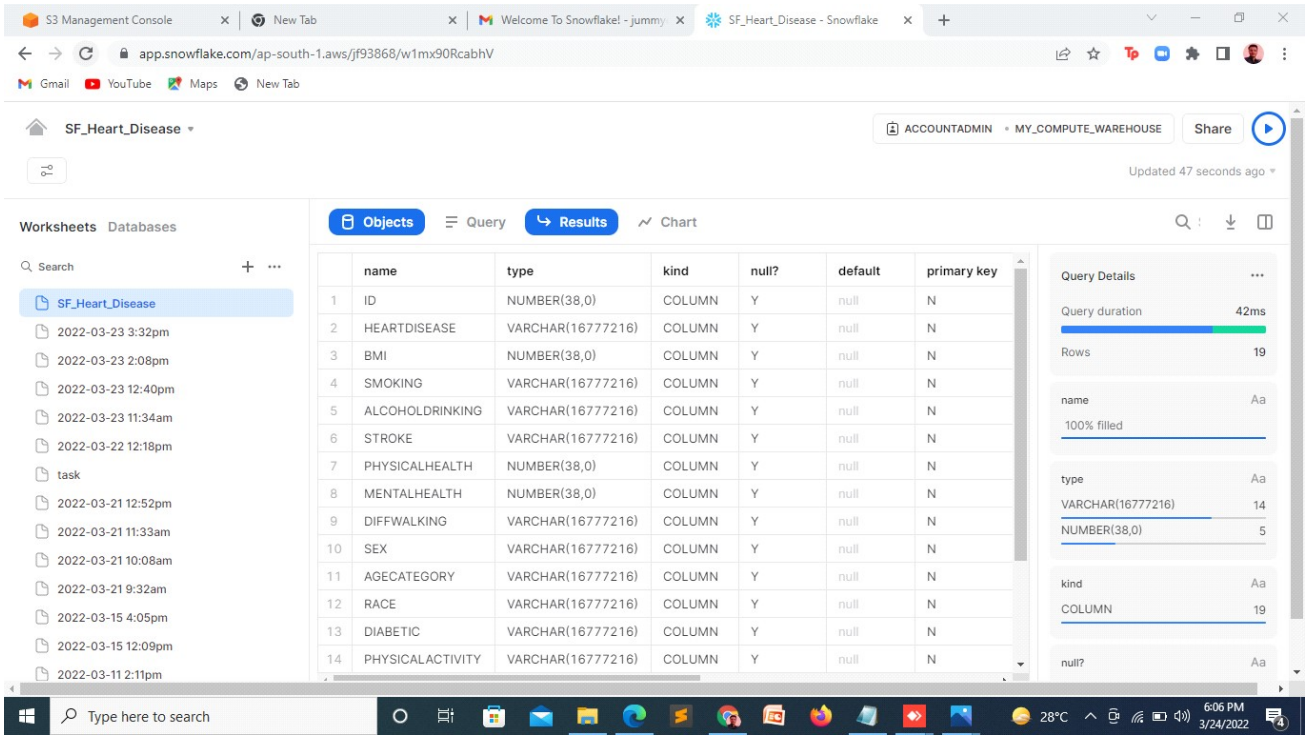


Fig 7.2

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.

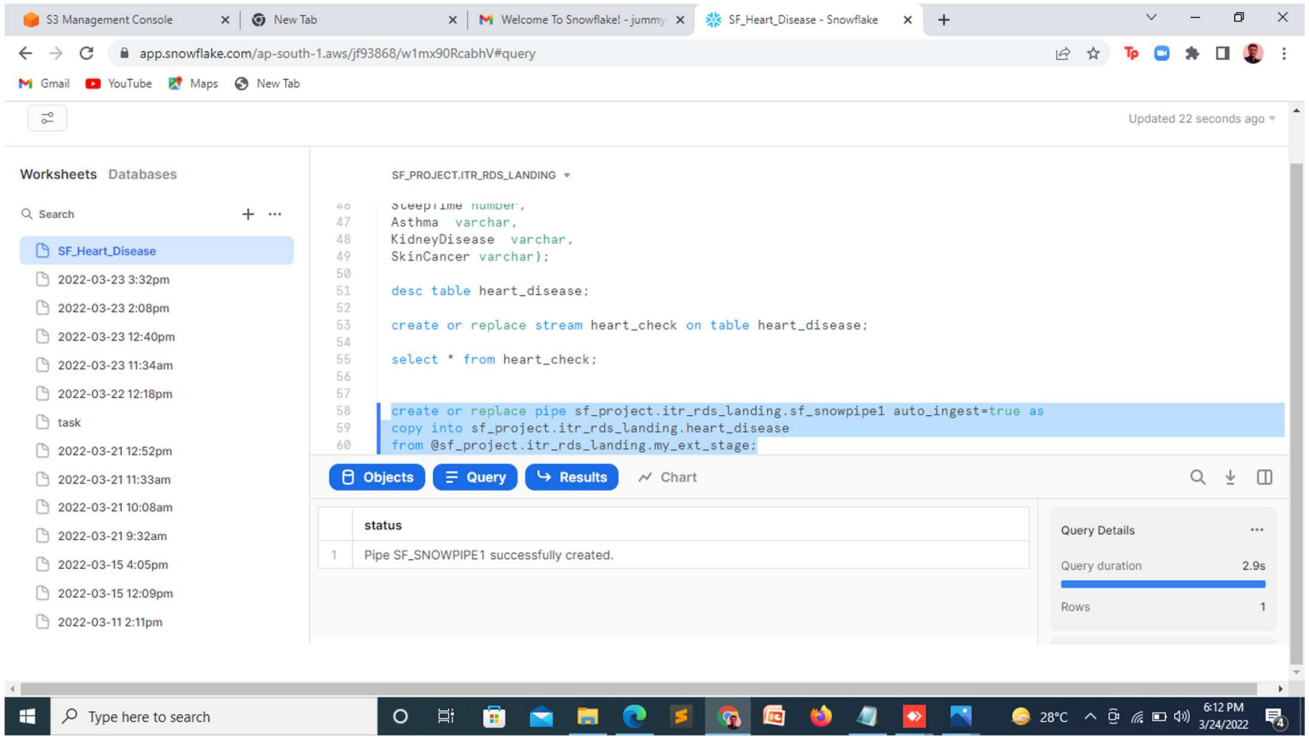


Fig 8

Query: show pipes;
Description: Showing the description of pipes created above.

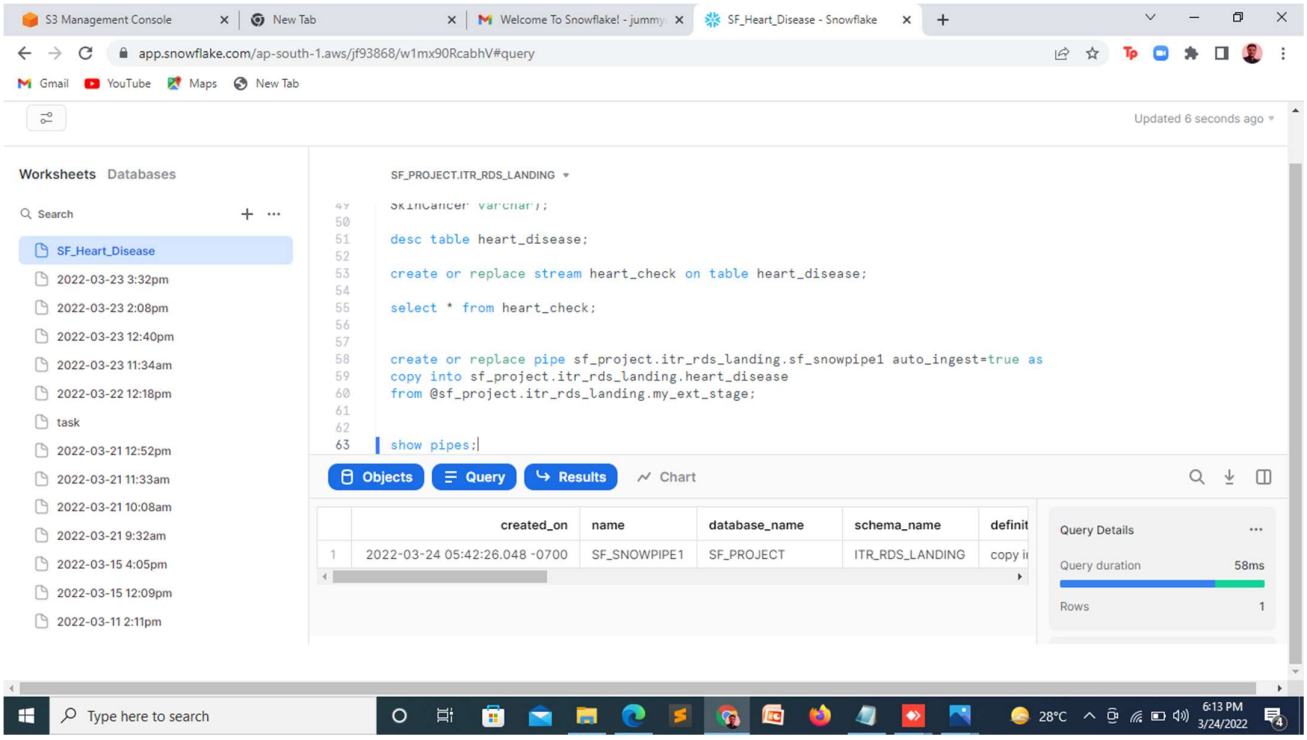


Fig. 9

Query: alter pipe sf_project.itr_rds_landing.sf_snowpipe1 refresh;
Description: Tells whether the dataset is sent or not.

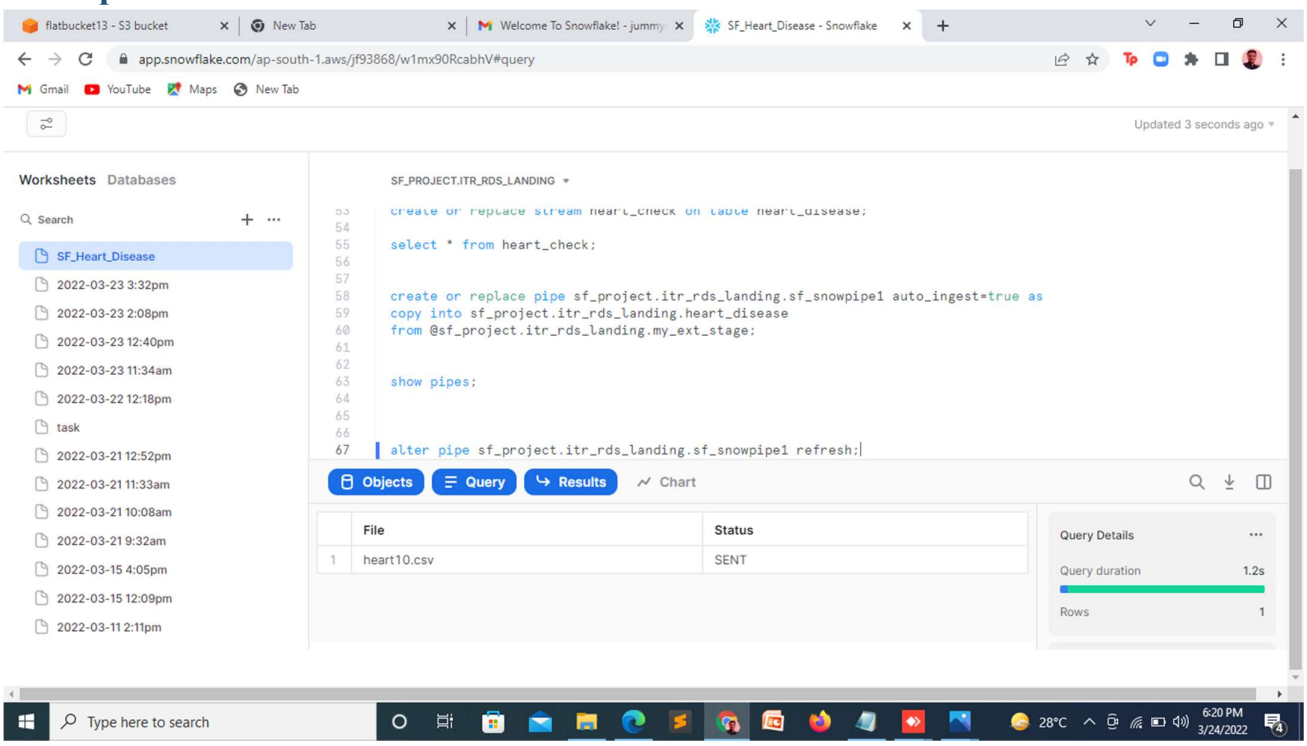


Fig. 10

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.

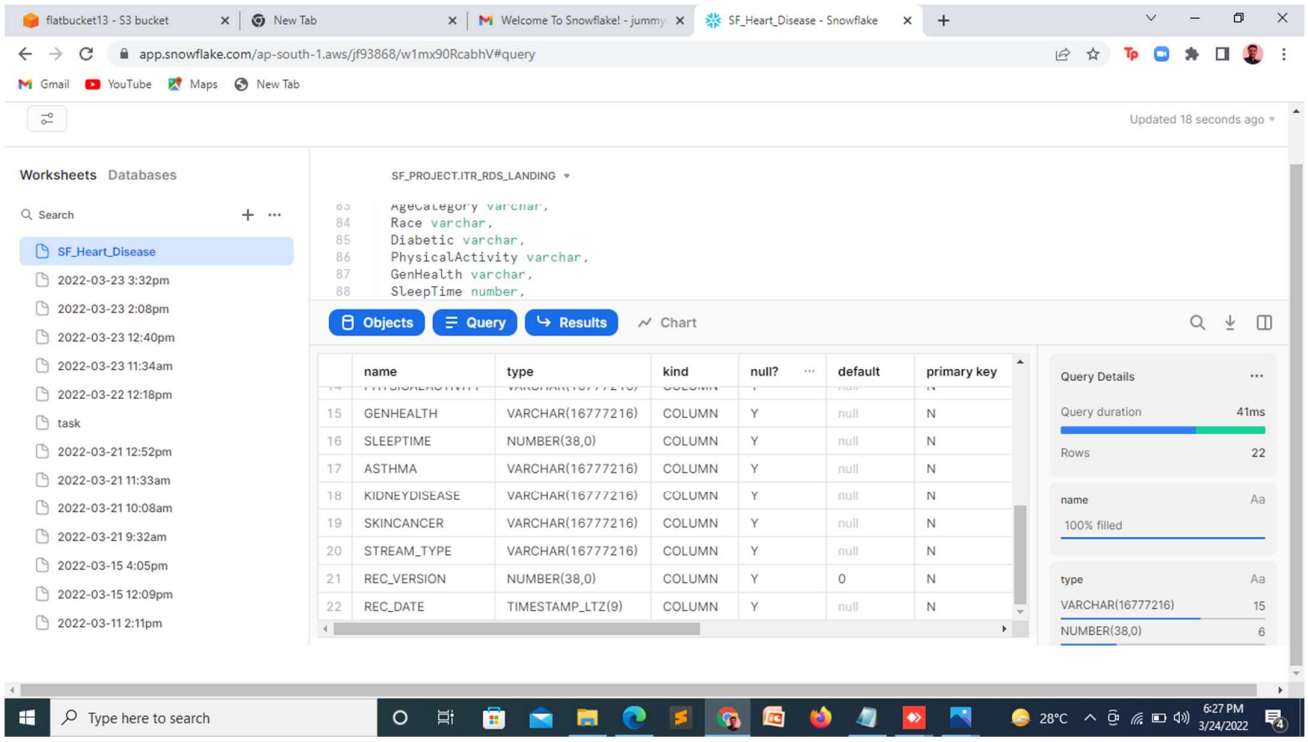


Fig. 11

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.

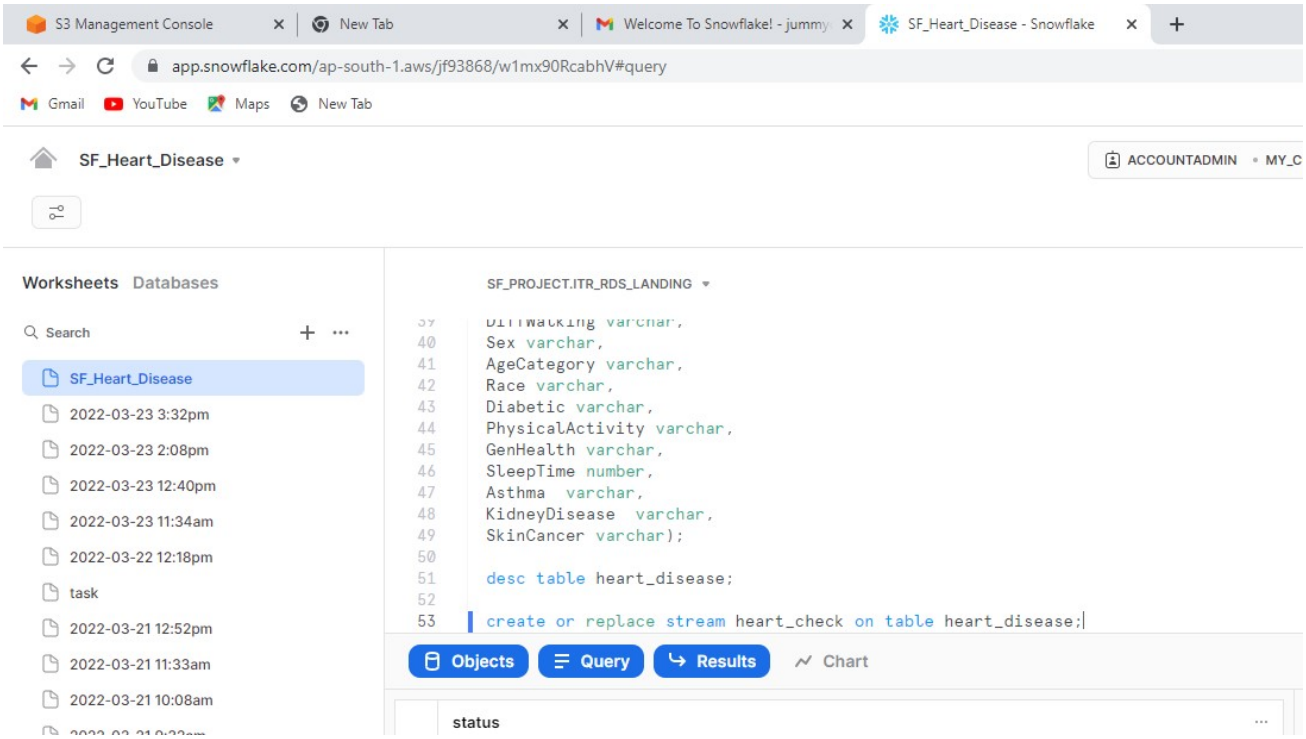


Fig. 12

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.

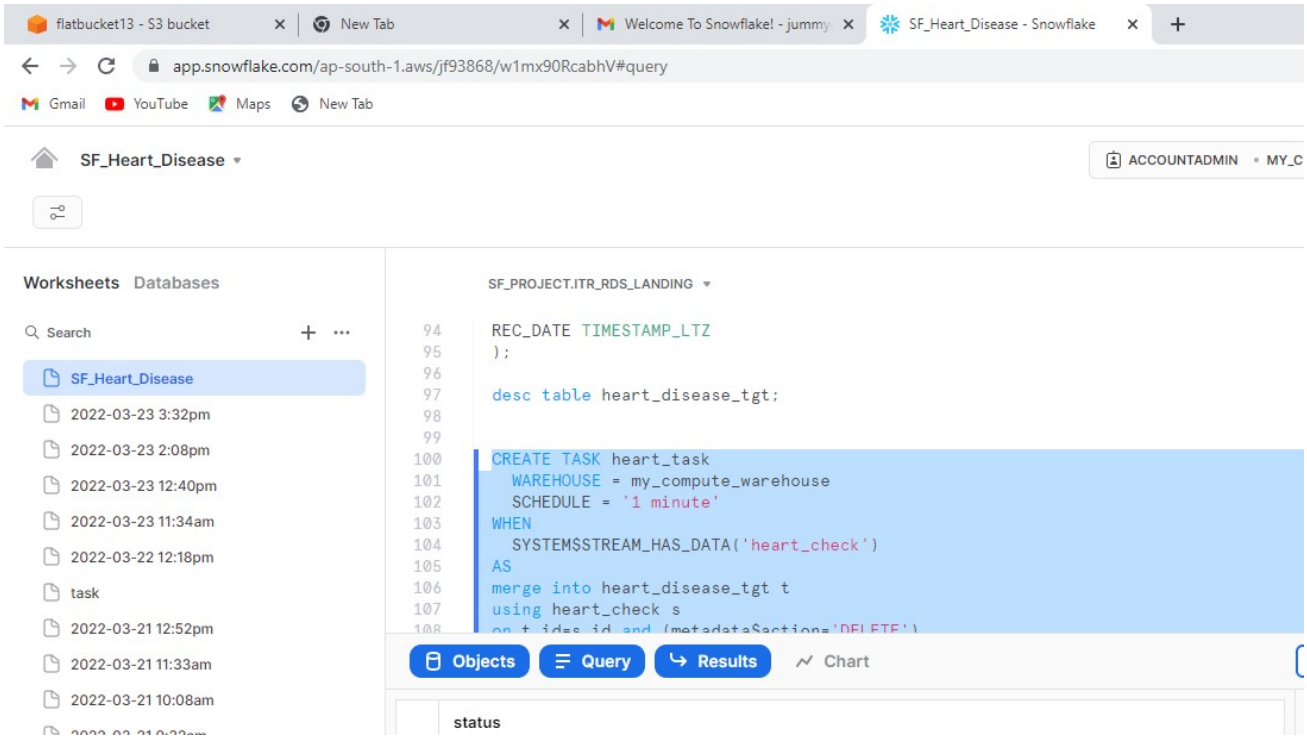


Fig. 13

Query: select * from heart_disease_tgt;
Description: showing the whole target table i.e HEART_DISEASE_TGT at once.

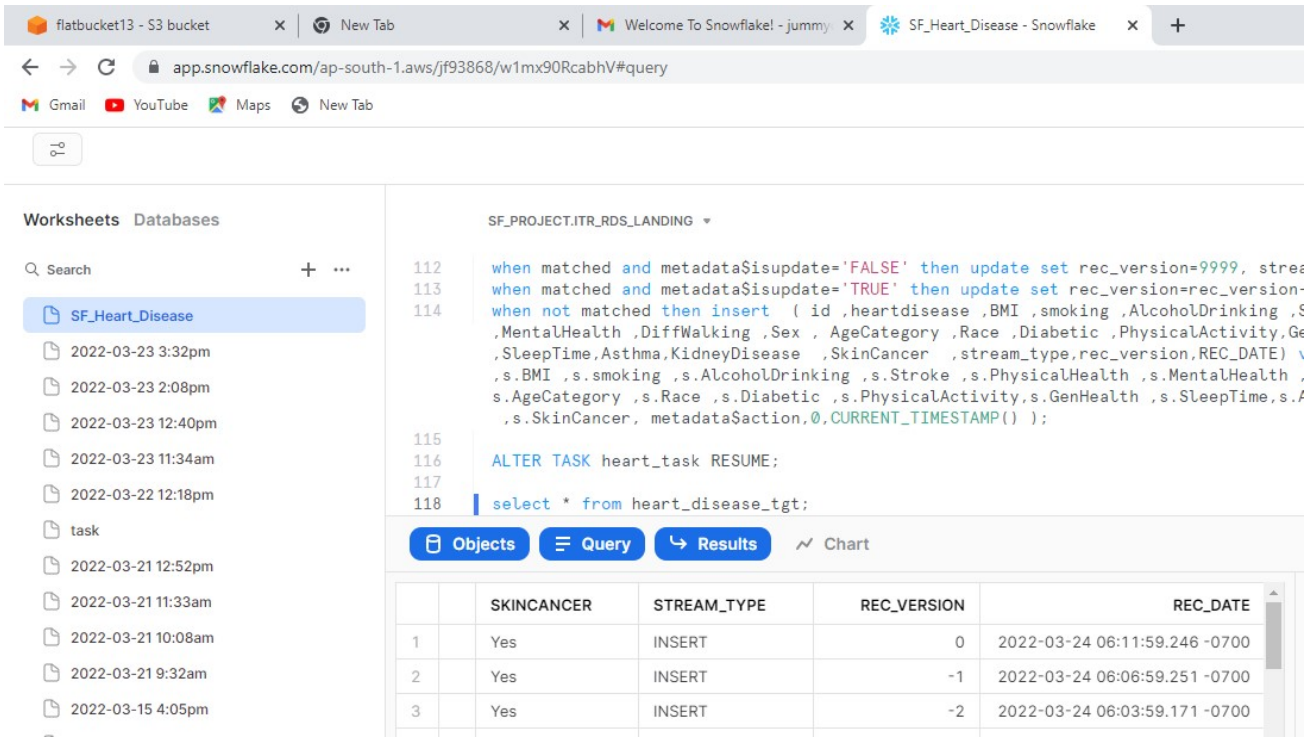
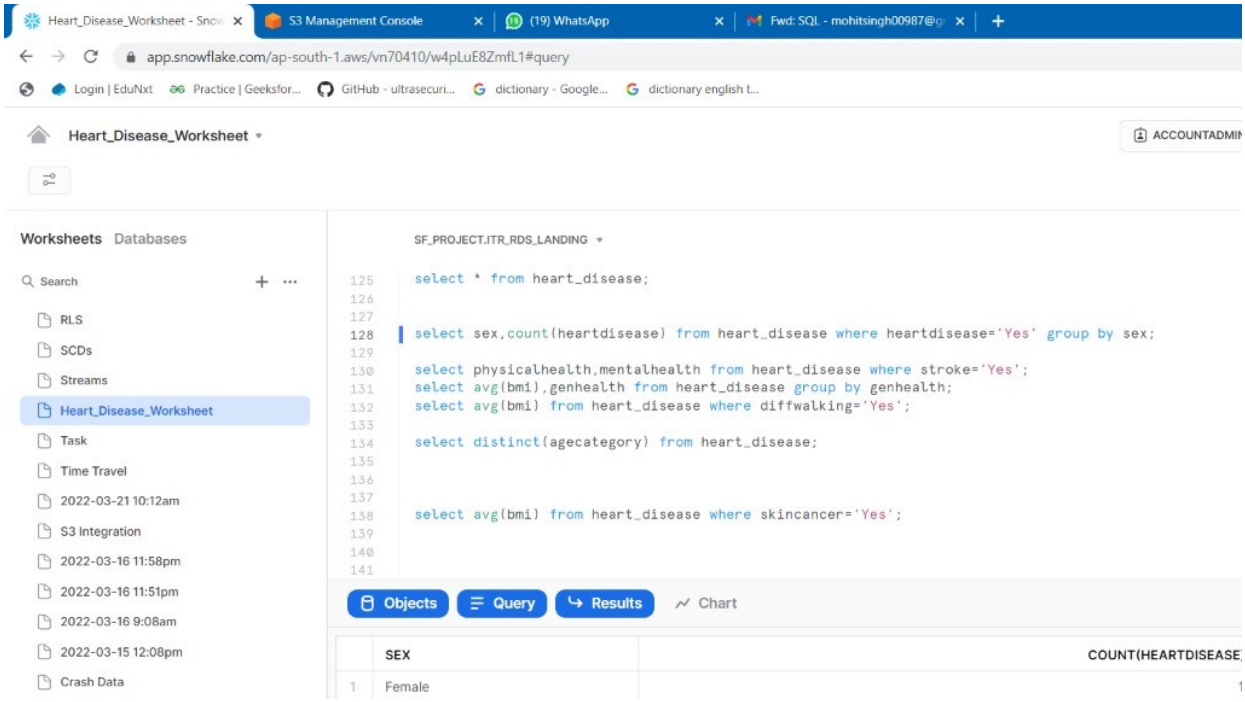


Fig. 14

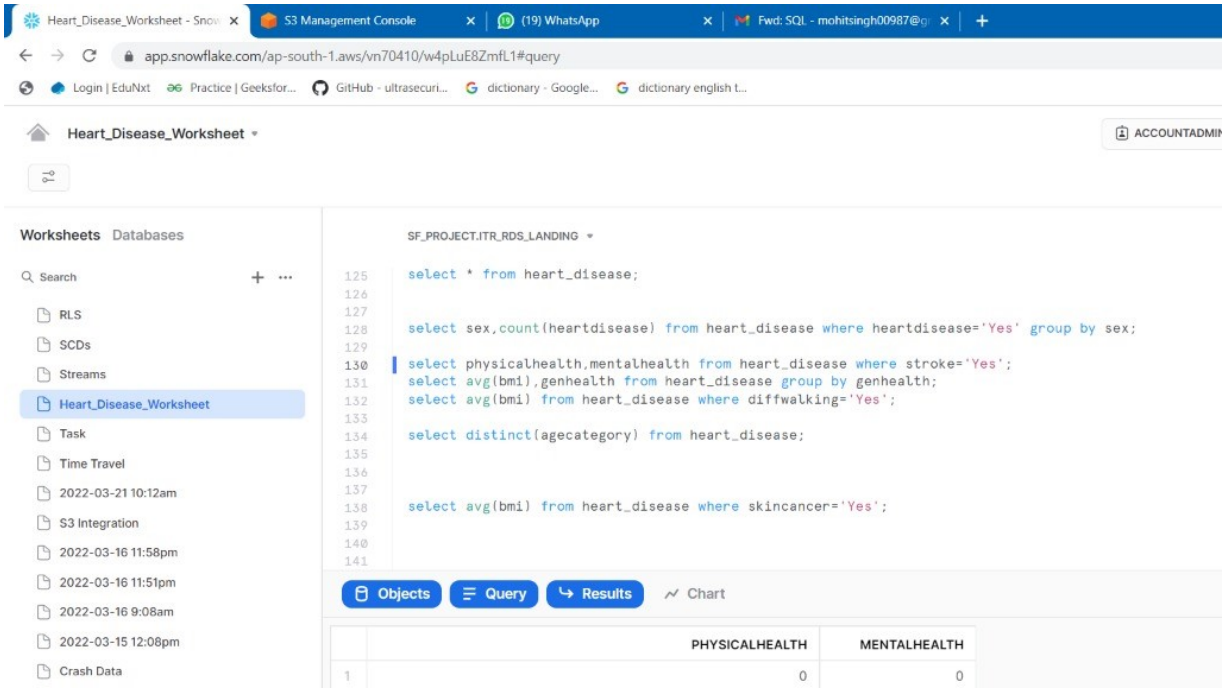
Data Analysis: -

- Which gender have majority heart attack?
  select sex, count(heartdisease) from heart_disease where heartdisease='Yes' group by sex;



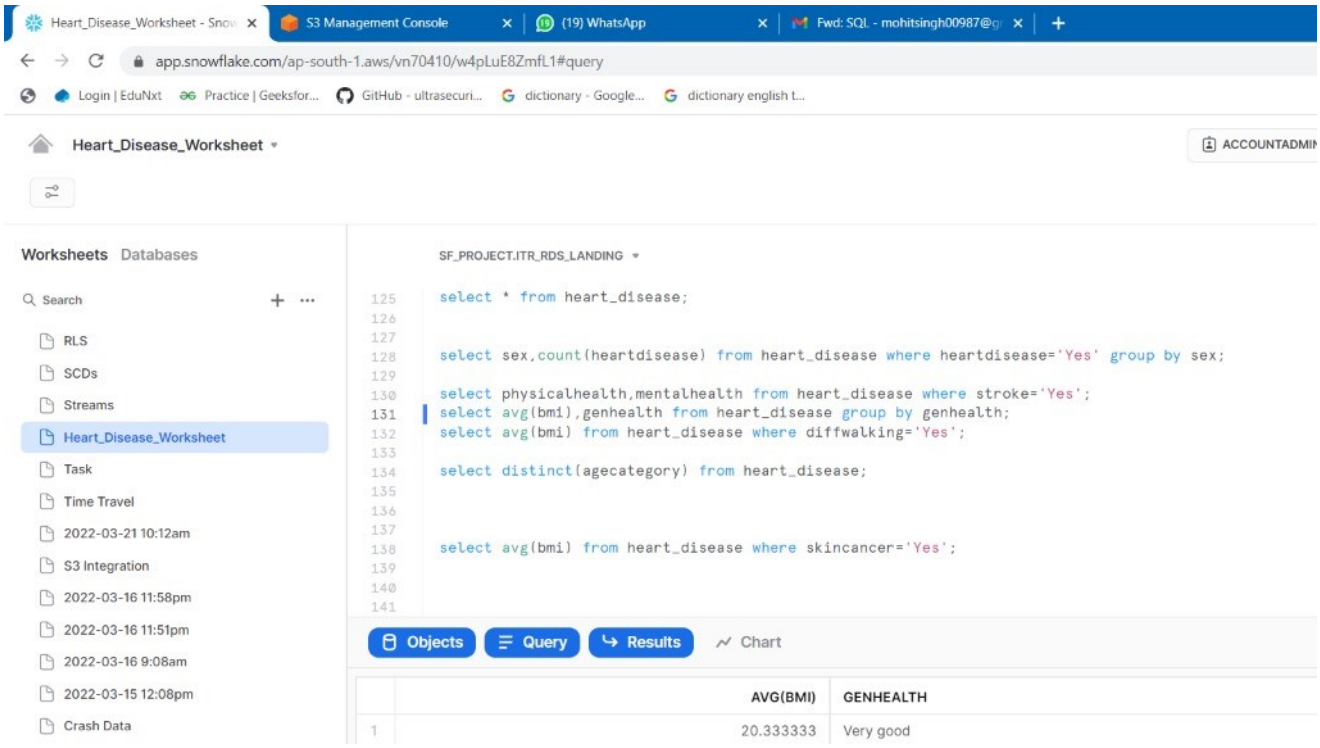
Solution 1

- Which combination of physical and mental health causes stroke?
  select physicalhealth, mentalhealth from heart_disease where stroke='Yes';



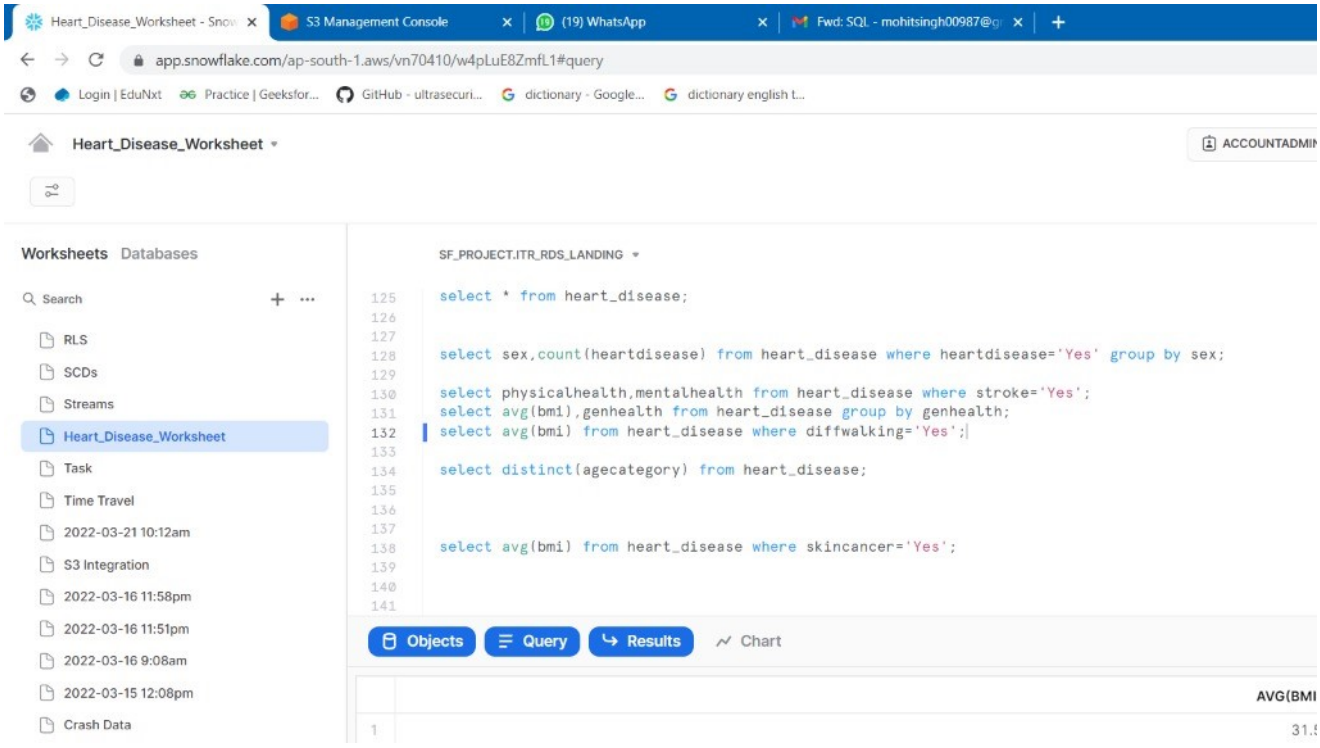
Solution 2

- GenHealth wise average BMI of people?
  select avg(bmi),genhealth from heart_disease group by genhealth;



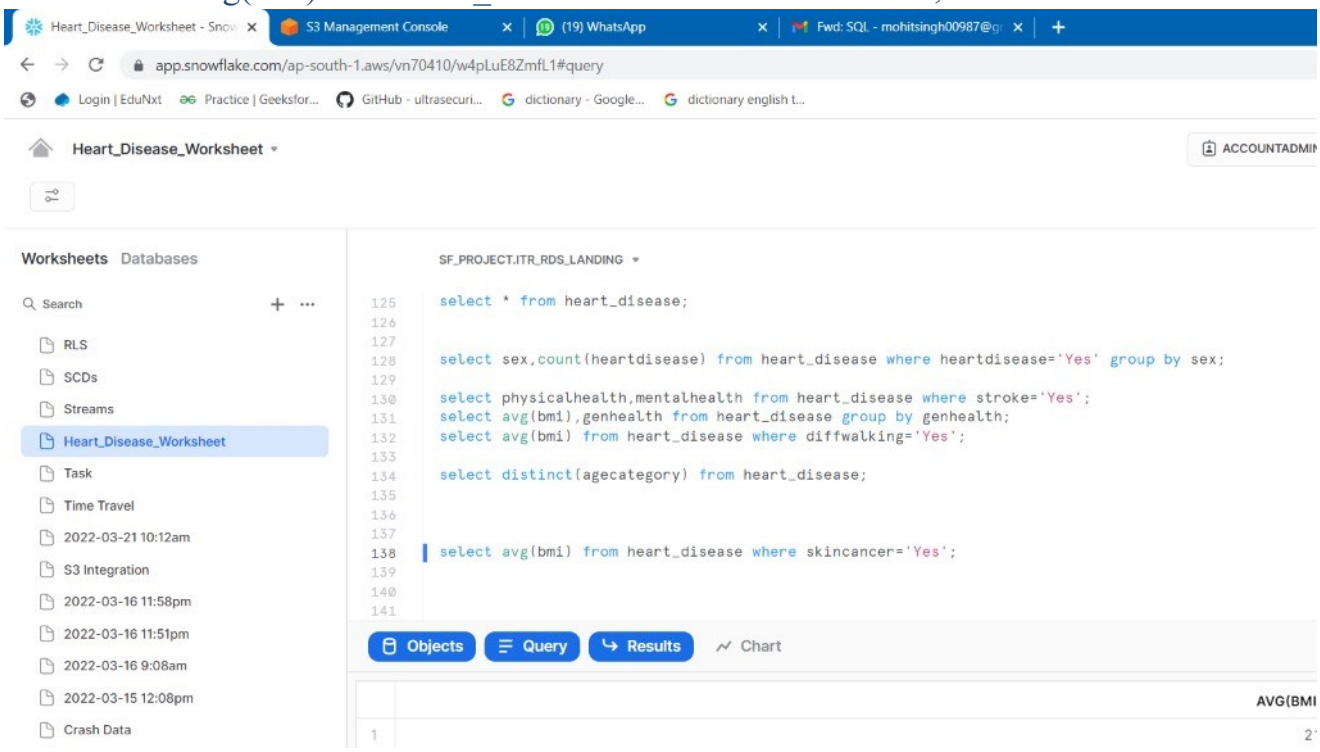
Solution 3

- Average age of people having problem in walking?
  select avg(bmi) from heart_disease where diffwalking='Yes';



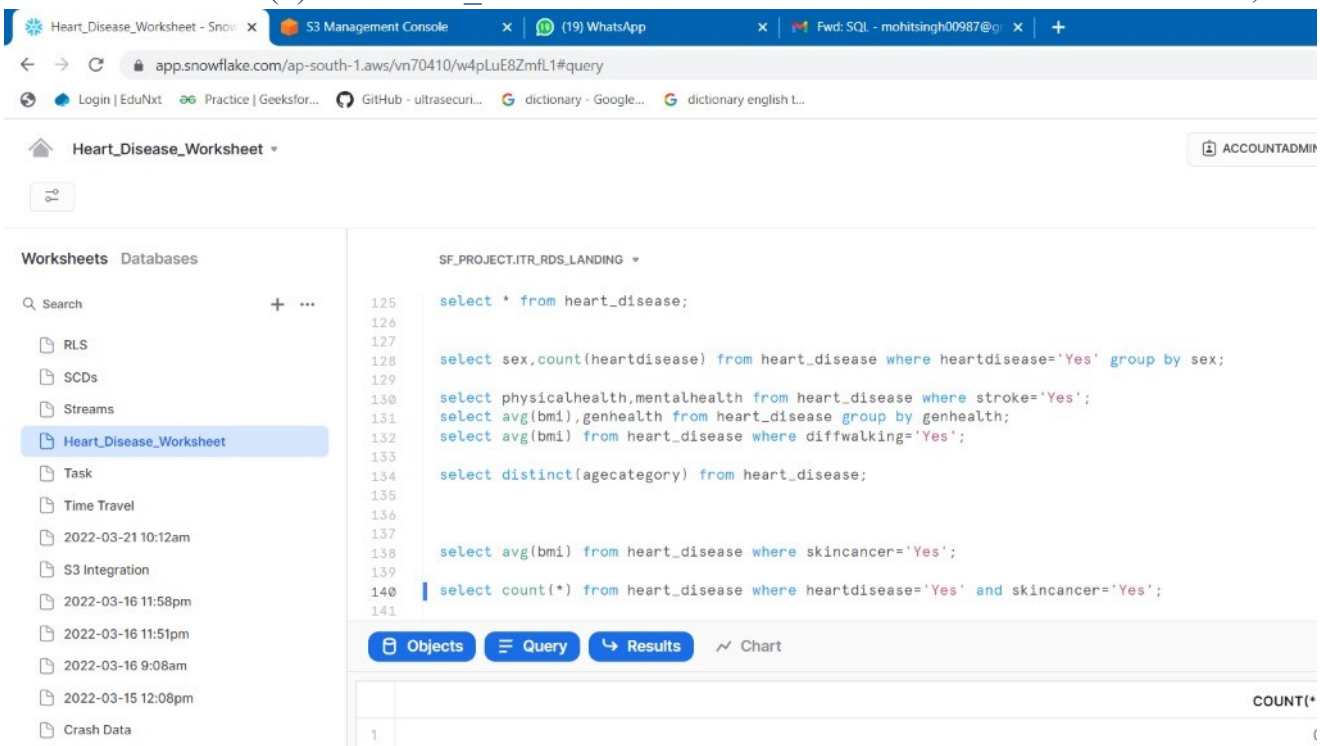
Solution 4

- 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