FINAL PROJECT on

DATA MIGRATION FROM TERADATA TO SNOWFLAKE

PREPARED BY:

ASSOCIATE ID	ASSOCIATE NAME
2272156	Subhankar Pandey
2273785	Subarna Paul
2272225	Unnimaya
2273807	Dhiraj Kumar
2272864	Rachit Alagundagi

GUIDED BY:

Mentor:

Kavin Kumar G

Associate

AIA-Cloud Data Integration

Coach:

Leon Benjamin

GenC HR Coach - DATA

Human Resource - GenC

CONTENTS

LIST OF ABREVIATIONS

TERM	DESCRIPTION
ETL	Extract, Transform and Load
VM	Virtual Machine
AWS	Amazon Web Services
S3	Simple Storage Services
CSV	Comma Separated Values
RDBMS	Relation Database System

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Decades ago, Teradata identified the need to manage and analyze large volumes of data. However, with the ever-changing landscape of data, including the exponential growth in Big Data the 3V's - volume, velocity, and variety, the cloud has emerged as a powerful enabler for modern data analytics. Taking advantage of cloud computing capabilities, Snowflake Cloud Data Platform has become a leading solution in the domain of cloud-based storage warehouses.

One of the significant advantages Snowflake offers over traditional systems like Teradata is its decoupling of compute and storage. This architecture allows Snowflake to develop a highly efficient cloud data platform that automatically and instantaneously scales both storage and compute capacity, a feat that was not feasible with Teradata, regardless of whether the Teradata system was on-premises or hosted in the cloud.

At the core of Snowflake's efficiency lies its multi-cluster, shared data architecture. This innovative approach ensures optimal resource utilization, enabling Snowflake to handle large workloads effectively and deliver high-performance data analytics.

With this migration from Teradata to Snowflake, businesses can unlock new possibilities for data-driven decision-making, empowering their teams with real-time analytics, advanced data modelling, and seamless integration with various data sources and business intelligence tools.

In summary, Snowflake's cloud-based storage warehouse represents a cutting-edge solution that addresses the limitations of traditional database systems like Teradata, and by embracing this modern data platform, organizations can propel their data analytics capabilities into a new era of efficiency and innovation.

2. Purpose of the Project

The goal of this project can be divided into two parts -

- The primary goal is to migrate our data from a legacy Teradata system to a cloud-native based system like Snowflake.
- Our secondary goal is to perform ETL process on this data

3. Functional Requirement 1: Hypothesis Generation

a. Problem Scenario

XYZ Company wants to migrate their employee details data from a legacy RDBMS to modern cloud-based data warehouse. They have collected the data, in Excel record format.

To perform the migration, they have chosen Snowflake as their data warehousing platform. The company wants to make a smooth transition of their user data with some additional transformation.

b. Problem Statement

As a data analyst at XYZ Company, your task is to perform the following steps:

- Data Loading and Exporting:
    - Create a table, 'EMP_DETAILS' to store the respective data.
    - Upload the 'EMP_DETAILS' CSV files to a Teradata database.
- Data Transformation:
    - Perform an ALTER operation on the 'EMP_DETAILS' table and add a new column named FULLNAME.
    - Perform a UPDATE operation on the 'EMP_DETAILS' table and add the full name of the employees under FULLNAME concatenating the first name and last name of the employees.
- Data Visualization:
    - Utilize Snowflake's dashboard feature to create visualizations and charts based on the analysis queries above.
    - Customize the dashboards to present key insights and trends.

By successfully completing this task, you will provide XYZ Company with valuable insights and visualizations to help them understand and monitor their employee details.

4. Functional Requirement 2: Data Exploration

a. Analysis & Requirement Gathering

Source:

The **EMP_DETAILS** dataset contains 1000 records of employee details with data fields like:

EMPLOYEE ID, FIRST_NAME, LAST_NAME, GENDER, JOB_TITLE, DEPARTMENT, SALARY, EMAIL.

SOURCE FILE NAME	DESCRIPTION	SOURCE FILE
EMP_DETAILS.csv	This is a comma delimited file with 8 input fields	 EMP_DETAILS.csv

Table 4.1: EMP_DETAILS dataset Source description

The top 3 records from the dataset are show in Fig 4.1:

```
EMPLOYEE_ID,FIRST_NAME,LAST_NAME,GENDER,JOB_TITLE,DEPARTMENT,SALARY,EMAIL
1,Kevina,Bugbee,Female,VP Quality Control,Research and
Development,109358,kbugbee0@businesswire.com
2,Gusty,Stannering,Polygender,Accountant II,Product Management,23486,gstannering1@mediafire.com
3,Asa,Shearmer,Male,VP Sales,Product Management,47553,ashearmer2@google.cn
```

Fig 4.1: Top 3 Sample Records from EMP_DETAILS dataset

5. Functional Requirement 3: Model Strategy

a. Designing Migration Workflow - Phase 1

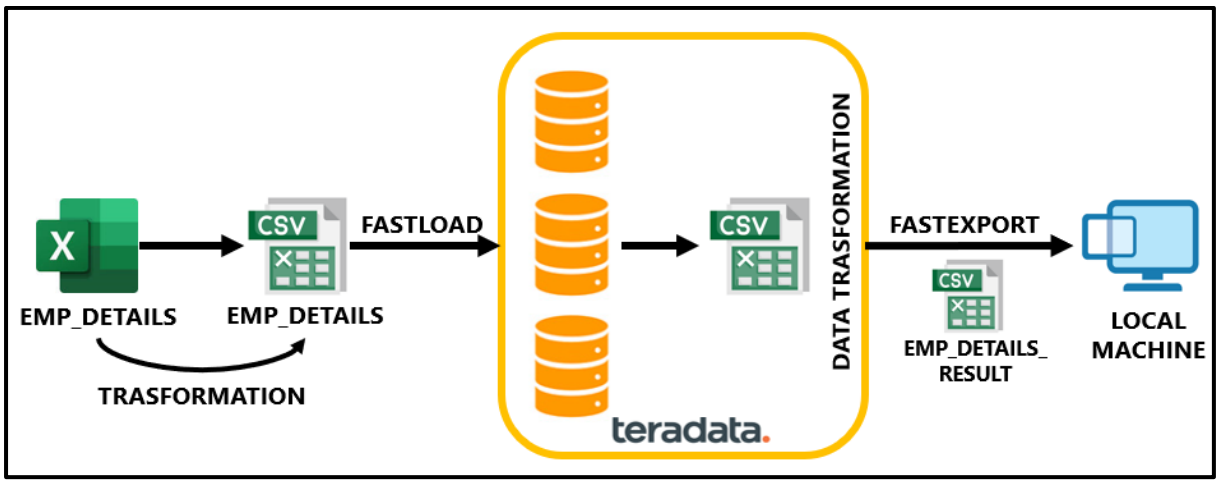


Fig 5.1: Source dataset to Teradata FastLoad and FastExport post Transformation

b. Designing Migration Workflow - Phase 2

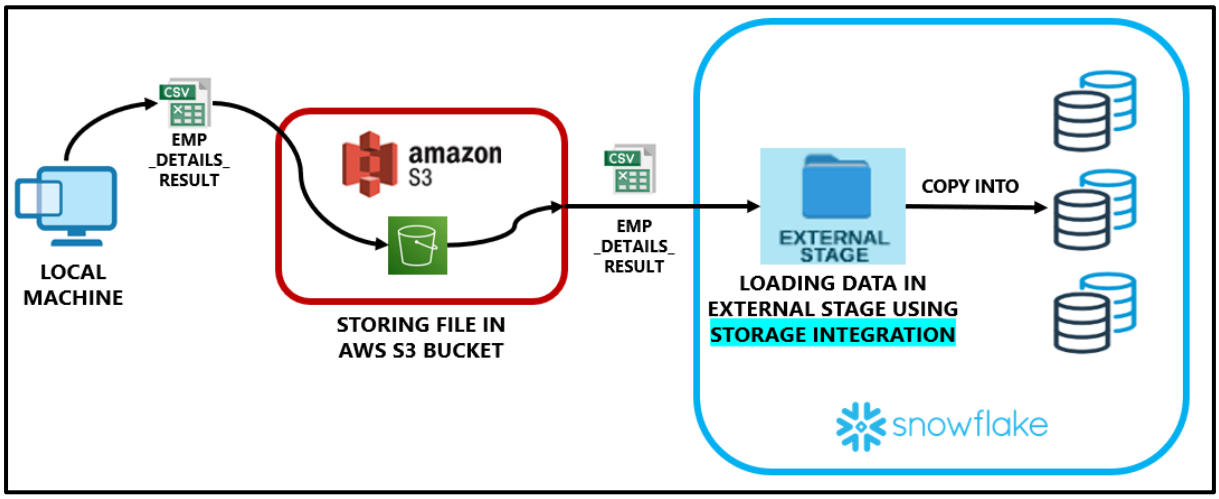


Fig 5.2: Host transformed dataset in AWS S3 bucket and establish connection with snowflake using Snowflake Storage Integration

6. Functional Requirement 4: Implementation & Coding

a. Performing ETL

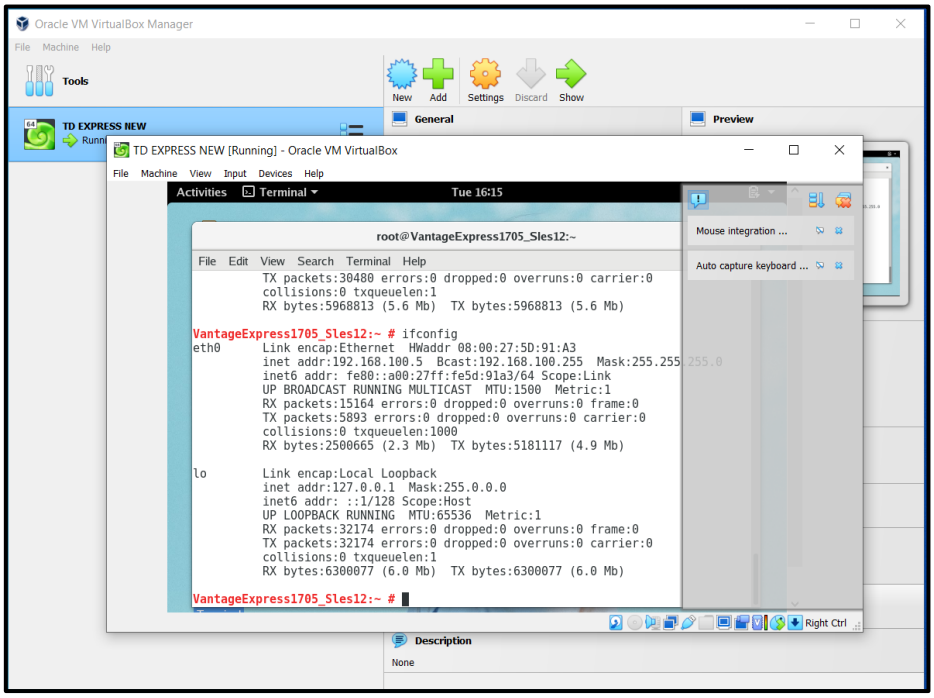


Fig 6.1: Launch the VM using Oracle Virtual Box and RUN the PDE or Parallel Database Extension to connect VM and Teradata Database

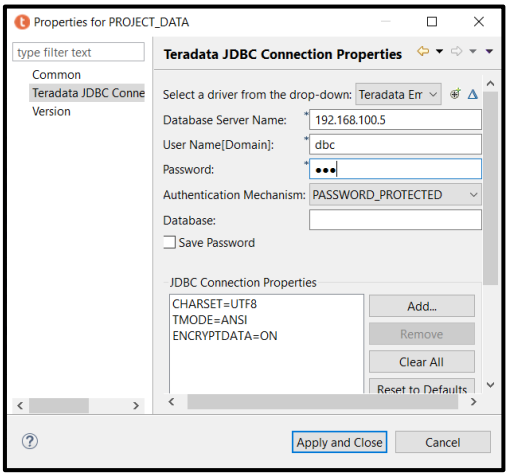


Fig 6.2: Insert Database Server Name as the inet address of the Virtual Machine

```
CREATE DATABASE EMPLOYEE
FROM "DBC" AS
PERM = 5368709120, -- 5GB
SPOOL = 5368709120, -- 5GB
TEMPORARY = 2147483648; -- 2GB

CREATE TABLE EMPLOYEE.DETAILS (
    EMPLOYEE_ID INTEGER NOT NULL,
    FIRST_NAME VARCHAR(100),
    LAST_NAME VARCHAR(100),
    GENDER VARCHAR(100),
    JOB_TITLE VARCHAR(100),
    DEPARTMENT VARCHAR(100),
    SALARY INTEGER,
    EMAIL VARCHAR(100),
    PRIMARY KEY(EMPLOYEE_ID)
)UNIQUE PRIMARY INDEX(EMPLOYEE_ID);
```

EMPLOYEE Database

DETAILS Table

Fig 6.3: Teradata DDL Commands to create a Database and Tables

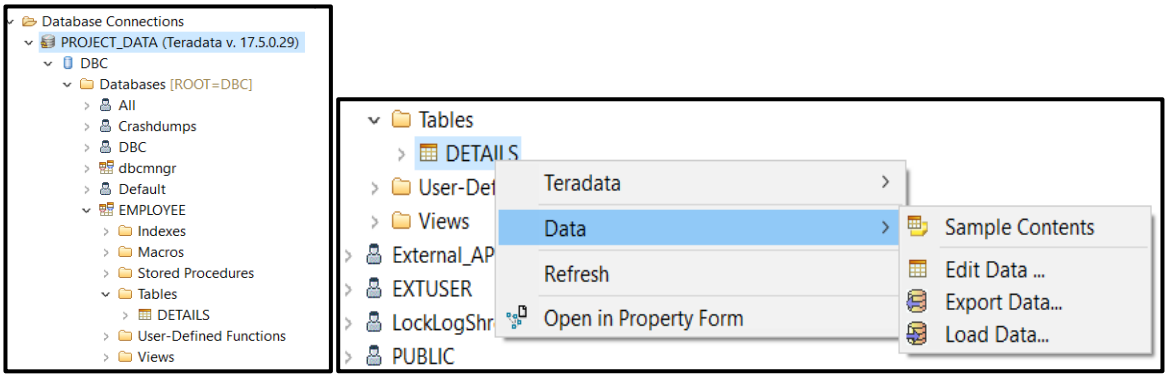
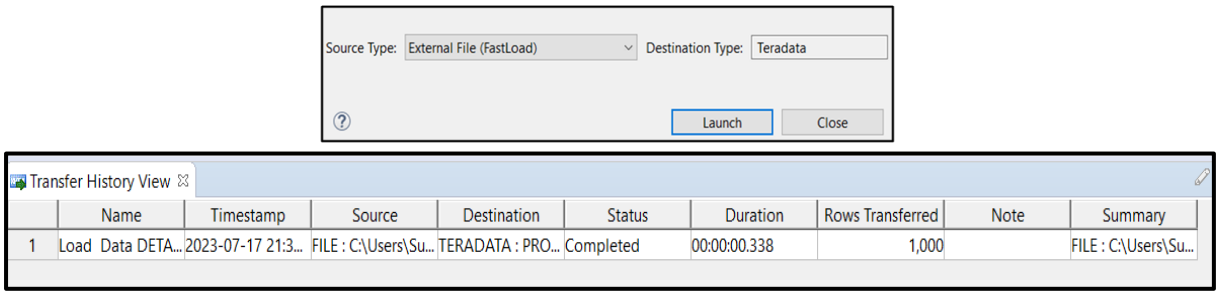


Fig 6.4: Perform data loading using Teradata FastLoad Utility



	Name	Timestamp	Source	Destination	Status	Duration	Rows Transferred	Note	Summary
1	Load Data DETA...	2023-07-17 21:3...	FILE : C:\Users\Su...	TERADATA : PRO...	Completed	00:00:00.338	1,000		FILE : C:\Users\Su...

Fig 6.5: Report of Loaded data in Transfer History View

```
ALTER TABLE EMPLOYEE.DETAILS
ADD FULLNAME VARCHAR(100);

UPDATE EMPLOYEE.DETAILS
SET FULLNAME = TRIM(FIRST_NAME) || ' ' || TRIM(LAST_NAME);

SELECT * FROM EMPLOYEE.DETAILS
ORDER BY EMPLOYEE_ID;
```

Fig 6.6: Data transformation, performs ALTER & UPDATE operation on the table

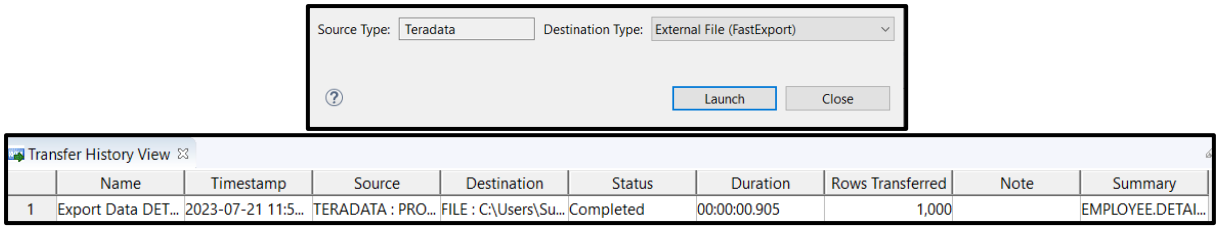


Fig 6.7: We are exporting the result data set, using Teradata FastExport

Target:

The **EMP_DETAILS_RESULT** dataset contains 1000 records of employee details with data fields like:

EMPLOYEE ID, FIRST_NAME, LAST_NAME, GENDER, JOB_TITLE, DEPARTMENT, SALARY, EMAIL, FULLNAME.

SOURCE FILE NAME	DESCRIPTION	SOURCE FILE
EMP_DETAILS_RESULT.csv	This is a comma delimited file with 9 input fields	EMP_DETAILS_RESULT.csv

Table 6.1: EMP_DETAILS_RESULT dataset Source description

b. Route via Cloud Storage: AWS S3

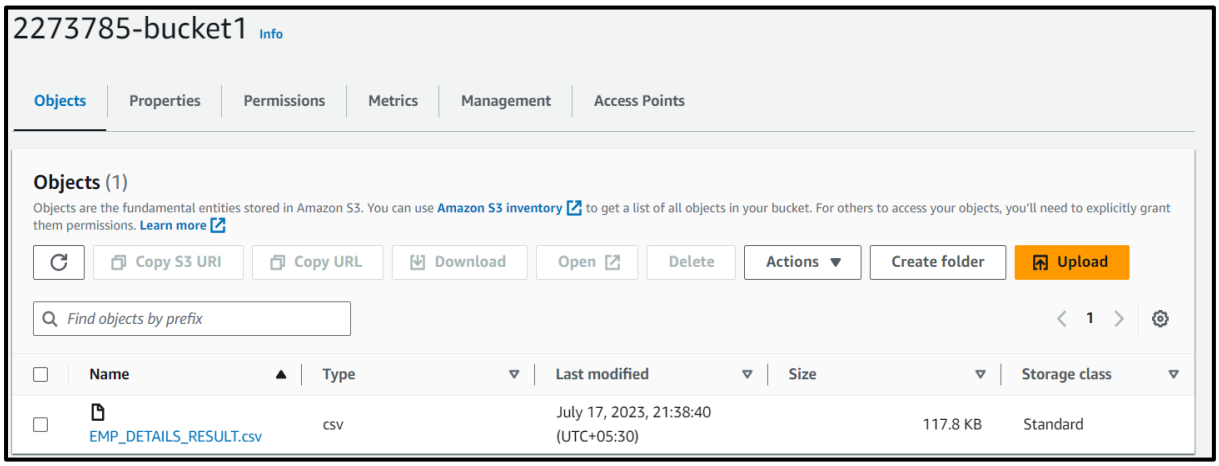


Fig 6.8: Transformed EMP_DETAILS_RESULT dataset will be stored inside 2273785-bucket1

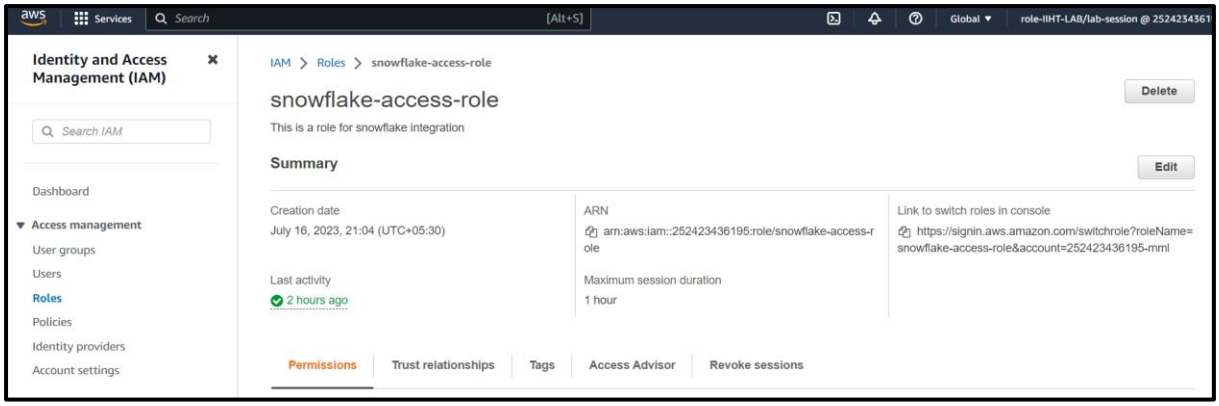


Fig 6.9: AWS Role ARN or Amazon Resource Number is used to create Integration Storage in Snowflake

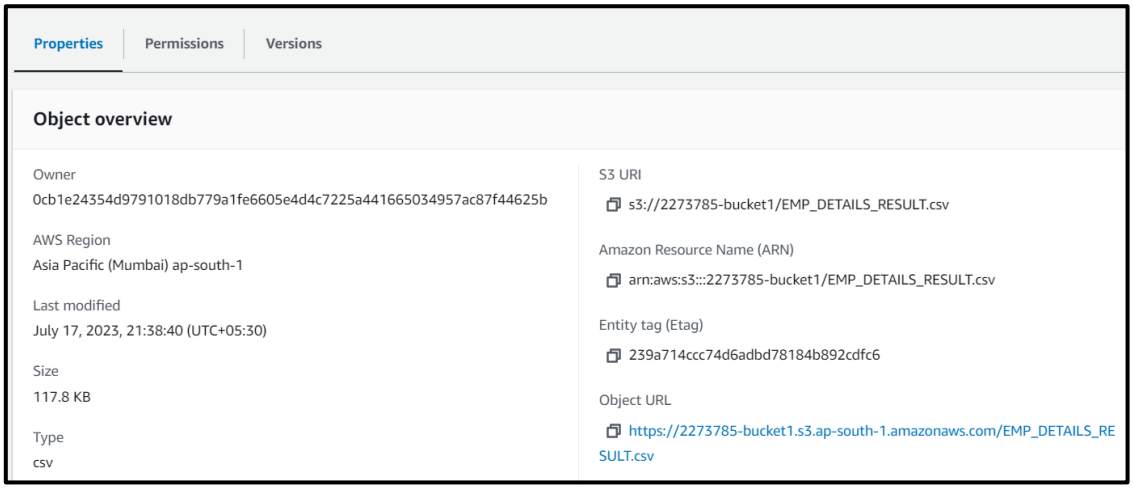


Fig 6.10: AWS S3 Bucket URI or Uniform Resource Identifier is used to locate the file storage in Snowflake

c. Snowflake Account Setup

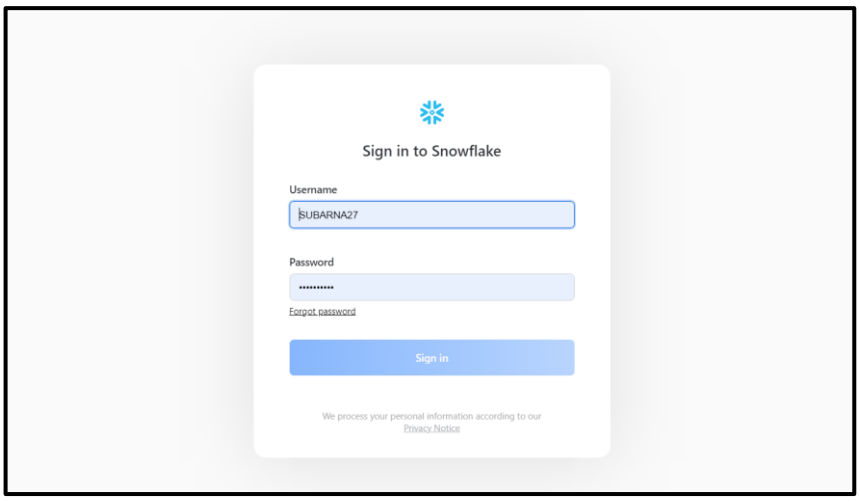


Fig 6.11 Snowflake account setup using email-id and password

d. DDL Script Migration

Teradata Data Types	Snowflake Data Types
INTEGER	INTEGER
VARCHAR (100)	VARCHAR (100)

Table 6.2: Mapping between Teradata and Snowflake data types

PROJECT.FINAL_PROJECT ▾ Settings ▾

```
-- WAREHOUSE CREATION
CREATE WAREHOUSE COMPUTE_WH WITH
WAREHOUSE_SIZE = 'SMALL'
WAREHOUSE_TYPE = 'STANDARD'
AUTO_SUSPEND = 600
AUTO_RESUME = TRUE
COMMENT = 'The Hands-On WH';

-- DATABASE CREATION
CREATE DATABASE PROJECT COMMENT = 'The Hands-On DB';
-- USING THE DATABASE
USE PROJECT;

-- SCHEMA CREATION
CREATE SCHEMA FINAL_PROJECT;

-- TABLE CREATION
CREATE OR REPLACE TABLE "PROJECT"."FINAL_PROJECT"."EMPLOYEE_DATA"(
EMPLOYEE_ID STRING,
FIRST_NAME VARCHAR(100),
LAST_NAME VARCHAR(100),
GENDER VARCHAR(100),
JOB_TITLE VARCHAR(100),
DEPARTMENT VARCHAR(100),
SALARY STRING,
EMAIL VARCHAR(100),
FULLNAME VARCHAR(100)
);
```

Fig 6.12: Migrating existing DDL Scripts to make them compatible with Snowflake DDL

e. Connect External Data Source

```
CREATE OR REPLACE STORAGE INTEGRATION AWSINT
TYPE = EXTERNAL_STAGE
STORAGE_PROVIDER = S3
ENABLED = TRUE
STORAGE_AWS_ROLE_ARN = 'arn:aws:iam::252423436195:role/snowflake-access-role'
STORAGE_ALLOWED_LOCATIONS = ('s3://2273785-bucket1/EMP_DETAILS_RESULT.csv')
COMMENT = 'This an optional comment'

DESC INTEGRATION AWSINT;
```

Fig 6.13: Creating a Storage in Snowflake with USER_ARN & EXTERNAL ID

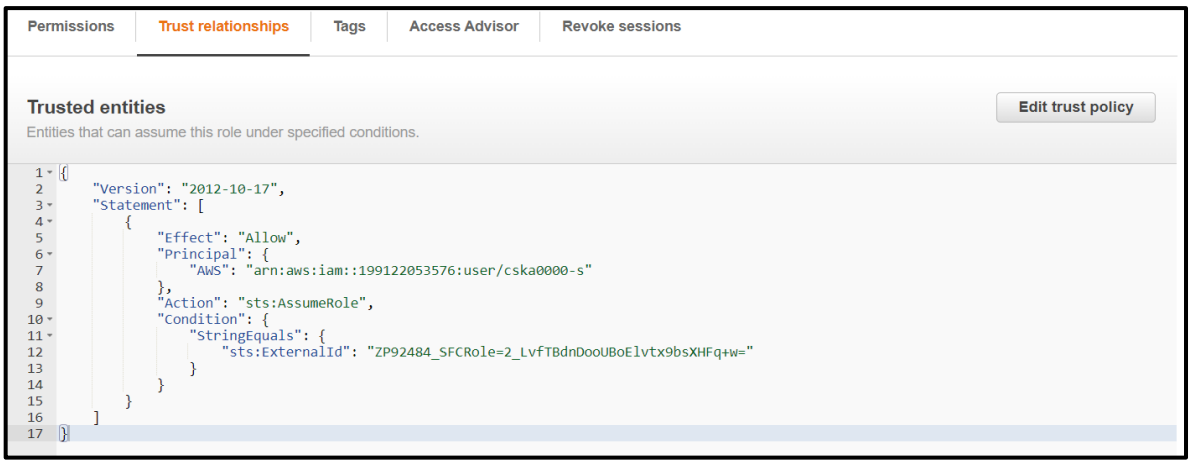
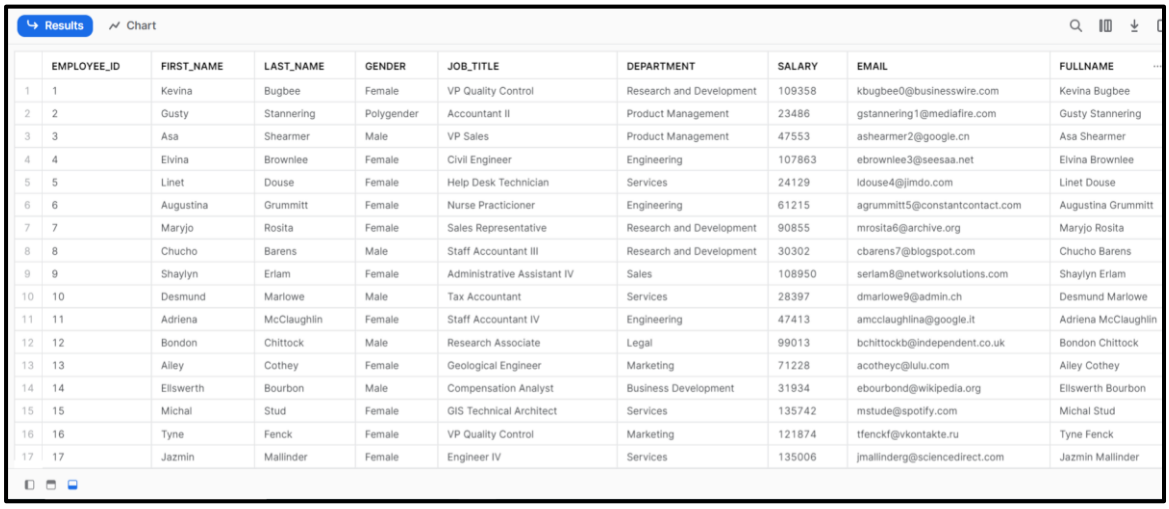


Fig 6.14: AWS custom Role Trust Policy is updated using the ARN & External ID generated by Snowflake Storage Integration

7. Testing & Validation

a. Migration Result:

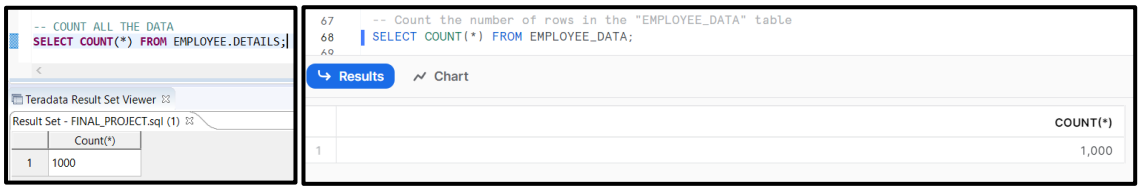


Results | Chart

	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	GENDER	JOB_TITLE	DEPARTMENT	SALARY	EMAIL	FULLNAME
1	1	Kevina	Bugbee	Female	VP Quality Control	Research and Development	109358	kbugbee0@businesswire.com	Kevina Bugbee
2	2	Gusty	Stannering	Polygender	Accountant II	Product Management	23486	gstannering1@mediafire.com	Gusty Stannering
3	3	Asa	Shearmer	Male	VP Sales	Product Management	47553	ashearmer2@google.cn	Asa Shearmer
4	4	Elvina	Browniee	Female	Civil Engineer	Engineering	107863	ebrowniee3@seesaa.net	Elvina Browniee
5	5	Linet	Douse	Female	Help Desk Technician	Services	24129	ldouse4@jimdo.com	Linet Douse
6	6	Augustina	Grummitt	Female	Nurse Practicioner	Engineering	61215	agrummitt5@constantcontact.com	Augustina Grummitt
7	7	Maryjo	Rosita	Female	Sales Representative	Research and Development	90855	mrosita6@archive.org	Maryjo Rosita
8	8	Chucho	Barens	Male	Staff Accountant III	Research and Development	30302	cbarens7@blogspot.com	Chucho Barens
9	9	Shaylyn	Erlam	Female	Administrative Assistant IV	Sales	108950	serlam8@networksolutions.com	Shaylyn Erlam
10	10	Desmund	Marlowe	Male	Tax Accountant	Services	28397	dmarlowe9@admin.ch	Desmund Marlowe
11	11	Adriena	McClaughlin	Female	Staff Accountant IV	Engineering	47413	amcclaughlina@google.it	Adriena McClaughlin
12	12	Bondon	Chittock	Male	Research Associate	Legal	99013	bchittockb@independent.co.uk	Bondon Chittock
13	13	Ailey	Cothey	Female	Geological Engineer	Marketing	71228	acotheyc@lulu.com	Ailey Cothey
14	14	Ellswerth	Bourbon	Male	Compensation Analyst	Business Development	31934	ebourbond@wikipedia.org	Ellswerth Bourbon
15	15	Michal	Stud	Female	GIS Technical Architect	Services	135742	mstude@spotify.com	Michal Stud
16	16	Tyne	Fenck	Female	VP Quality Control	Marketing	121874	tfenckf@vkontakte.ru	Tyne Fenck
17	17	Jazmin	Mallinder	Female	Engineer IV	Services	135006	jmallinderg@sciencedirect.com	Jazmin Mallinder

Fig 7.1: EMP_DETAILS is successfully migrated from Teradata to Snowflake

b. Integration Testing:



```
-- COUNT ALL THE DATA
SELECT COUNT(*) FROM EMPLOYEE.DETAILS;
```

Teradata Result Set Viewer

Result Set - FINAL_PROJECT.sql (1)

	Count(*)
1	1000

```
67  -- Count the number of rows in the "EMPLOYEE_DATA" table
68  SELECT COUNT(*) FROM EMPLOYEE_DATA;
```

Results | Chart

	COUNT(*)
1	1,000

Fig 7.2: COUNT of Records in Teradata equals with Snowflake Records COUNT

c. Data Transformation Testing:

SR. NO.	TABLE	REQUIREMENT	DESCRIPTION	INPUT	EXCPECTED OUTPUT	OUTPUT	STATUS
1	EMPLOYEE_DATA	Integration Testing	Check the intregation between Teradata and Snowflake	LIST @SF_ST;	s3://2273785-bucket1/csv/EMP_DETAILS_RESULT.csv	s3://2273785-bucket1/csv/EMP_DETAILS_RESULT.csv	PASSED
2	EMPLOYEE_DATA	Testing specific column values	Validate the column names and data types of the "EMPLOYEE_DATA" table	DESCRIBE TABLE EMPLOYEE_DATA;	List out all the column names	List out all the column names	PASSED
3	EMPLOYEE_DATA	Checking data uniqueness	Test for data uniqueness	SELECT COUNT(DISTINCT EMPLOYEE_ID) AS EMPLOYEE FROM EMPLOYEE_DATA;	1000	1000	PASSED
4	EMPLOYEE_DATA	Testing for NULL values	Check if there is any NULL value exists in the transformed dataset	SELECT * FROM EMPLOYEE_DATA WHERE FIRST_NAME IS NULL;	0	0	PASSED
5	EMPLOYEE_DATA	Performing data aggregation	Calculate total amount of Sales based on the Product Category	SELECT DEPARTMENT, AVG(SALARY) AS AvgSalary FROM EMPLOYEE_DATA GROUP BY DEPARTMENT;	Research and Development 84,815.860465116 Product Management 83,207.038461538 Engineering 83,648.012658228	Research and Development 84,815.860465116 Product Management 83,207.038461538 Engineering 83,648.012658228	PASSED
6	EMPLOYEE_DATA	Verifying data transformations	Count all the employee's first name, last name and full name	SELECT COUNT(FIRST_NAME) AS Total_FirstName, COUNT(LAST_NAME) AS Total_LastName, COUNT(FULLNAME) AS Total_FullName FROM EMPLOYEE_DATA;	TOTAL_FIRSTNAME 1000 TOTAL_LASTNAME 1000 TOTAL_FULLNAME 1000	TOTAL_FIRSTNAME 1000 TOTAL_LASTNAME 1000 TOTAL_FULLNAME 1000	PASSED

d. Data Visualization Dashboard:

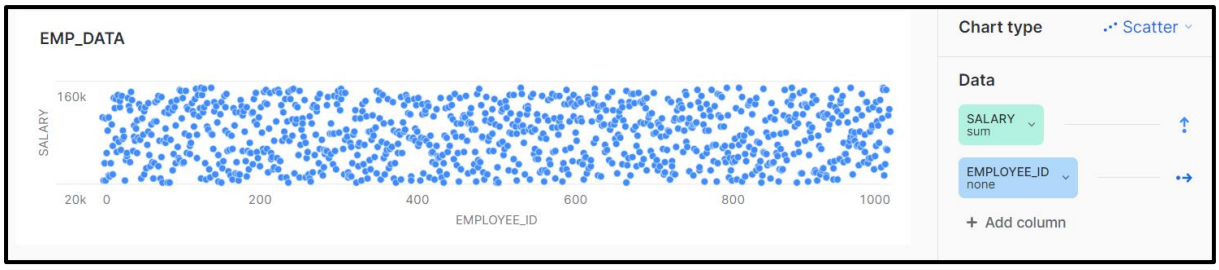


Fig 7.3: DASHBOARD is created to visualize insights from the employee data

8. Conclusion

We have completed Teradata to Snowflake migration.

By successfully completing this project, we have:

Data Migration: Successfully moved our Data Model from legacy RDBMS (Teradata) to a modern Cloud-Native Data Warehouse (Snowflake).

DDL Migration: Migrated existing DDL scripts and developed new ones for seamless transition.

Data Transformation: Performed complex data transformations to consolidate multiple tables into a single virtual table.

Data Loading: Efficiently loaded data from Teradata to Cloud Storage (AWS S3) for streamlined access.

Integration Setup: Connected Snowflake to our Data Source via Storage Integration object for real-time data updates.

Quality Testing: Conducted scenario-based testing to ensure data quality and accurate transformation.

Decommissioning: Successfully decommissioned Teradata, providing valuable insights and visualizations for employee details monitoring.

9. Acknowledgement:

We would like to express our sincere gratitude and appreciation to our mentor & trainer **Mr. Kavin Kumar Govindaraj** for his invaluable support throughout this project.

We are truly grateful for his guidance, constant encouragement, valuable suggestions, innovative ideas, mentorship and supervision throughout this project work, which have significantly enriched our learning experience and helped us to complete the project successfully.