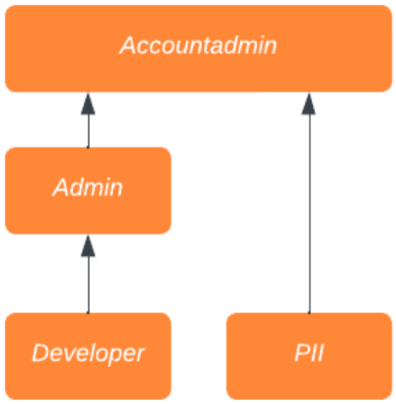
SNOWFLAKE ASSIGNMENT

PEER LEARNING DOCUMENT

Problem Statement -

1. Create roles as per the below-mentioned hierarchy. Accountadmin already exists in Snowflake (10).



2. Create an M-sized warehouse using the accountadmin role, name -> assignment_wh and use it for all the queries (5).
3. Switch to the admin role (5).
4. Create a database assignment_db (5)
5. Create a schema my_schema (5)
6. Create a table using any sample csv. You can get 1 by googling for sample csv's. Preferably search for a sample employee dataset so that you have PII related columns else you can consider any column as PII (5).
7. Also, create a variant version of this dataset (5).
8. Load the file into an external and internal stage (5).
9. Load data into the tables using copy into statements. In one table load from the internal stage and in another from the external (10).
10. Upload any parquet file to the stage location and infer the schema ofthefile(5).
11.Run a select query on the staged parquet file without loading it to a snowflake table (5).
12. Add masking policy to the PII columns such that fields like email, phone number, etc. show as **masked** to a user with the developer role. If the role is PII the value of these columns should be visible (15).

Amit Shukla

Approach -

He had written a script is written in SQL language that is used to perform given queries in Snowflake. Below is the explanation of each query -

1. He created three roles:
    - Admin
    - Developer
    - PII
2. He granted the Admin role to the ACCOUNTADMIN role, the Developer role to the Admin role, and the PII role to the ACCOUNTADMIN role.
3. He created a medium-sized data warehouse using the ACCOUNTADMIN role and grants all privileges on the warehouse to the Admin role.
4. He created a database named "assignment_db" and a schema named "my_schema" inside the database, using the Admin role.
5. He created a table named "EMPLOYEES_USING_EXT_STAGE" with columns for first name, last name, email, location, department, ELT timestamp, ELT by, and file name.
6. He created an external stage named "my_ext_stage" that points to an S3 bucket and allows the Admin role to use it.
7. He created a file format named "my_file_format" that specifies CSV format with a comma as the field delimiter and header row skipped.
8. He loaded data from the external stage into a table named "EMPLOYEES_JSON" that has a single column named "emp_data" of type VARIANT.
9. He created a table named "EMPLOYEES_VARIANT" that extracts data from the "emp_data" column and maps it to columns with appropriate data types.
10. He loaded data from an internal stage named "EMPLOYEES_USING_INT_STAGE" into a table named "EMPLOYEES_USING_INT_STAGE" using the COPY INTO command.
11. He created a file format named "my_parquet_ff" that specifies Parquet format.
12. He created a stage named "my_parquet_stage" that points to an S3 bucket with Parquet files.
13. He infered the schema of the Parquet file using the INFER_SCHEMA function and selects data from the stage.
14. It creates a masking policy named "email_mask" that masks the email column of the "EMPLOYEES_USING_EXT_STAGE" table.
15. He altered the "EMAIL" column of the "EMPLOYEES_EXT_STAGE" table to apply the masking policy.
16. He granted usage on the warehouse to the Developer role and grants usage on the database and schema to the PII role using the ACCOUNTADMIN role to test masking policy.

Rahul Kumar

Approach -

He had written a script is written in SQL language that is used to perform given queries in Snowflake. Below is the explanation of each query -

1. He created three roles:
    a. Admin
    b. Developer
    c. PII
2. He granted the admin role to accountadmin and the developer role to admin
3. He created a medium-sized data warehouse called assignment_wh
4. He granted all privileges on the data warehouse to the admin role and CREATE DATABASE on the account to the admin role
5. He created a database called assignment_db and a schema called my_schema
6. He created three tables: employee, in_employee, and json_employee
7. He created a file format called json_format for JSON files
8. He created a stage called json_stage for JSON files
9. He copied data from a JSON file into the json_employee table
10. He created a stage called internal_stage for CSV files
11. He copied data from a CSV file in the internal_stage to the in_employee table
12. He created a storage integration object for an S3 bucket, grants all privileges to the admin role, and describes the integration object
13. He created a file format called my_csv_format for CSV files
14. He creates a stage called external_stage for CSV files in the S3 bucket
15. He copied data from the external_stage to the employee table
16. He created a stage called parquet_stage for Parquet files
17. He infered the schema of a Parquet file in the parquet_stage
18. He queried the Parquet file from the parquet_stage
19. Hr granted privileges to the developer role to use the database, schema, and table in_employee and to select from it
20. He created a masking policy to hide email on the in_employee table for the developer role