Name: Shubham Manisha Naik **SJSU ID:**017627025
Name: Shreyas Vinayak Mohite **SJSU ID:**018207475
Name: Nitya Rondla **SJSU ID:**018204186
Name: Rutuja Nitin Kadam **SJSU ID:**018207176

1] Set up the development environment:

- Install Python 3.8+ on your system (Mac)

```
(base) shubhamnaik@Shubhams-MacBook-Air connection % python3 --version
Python 3.13.0
```

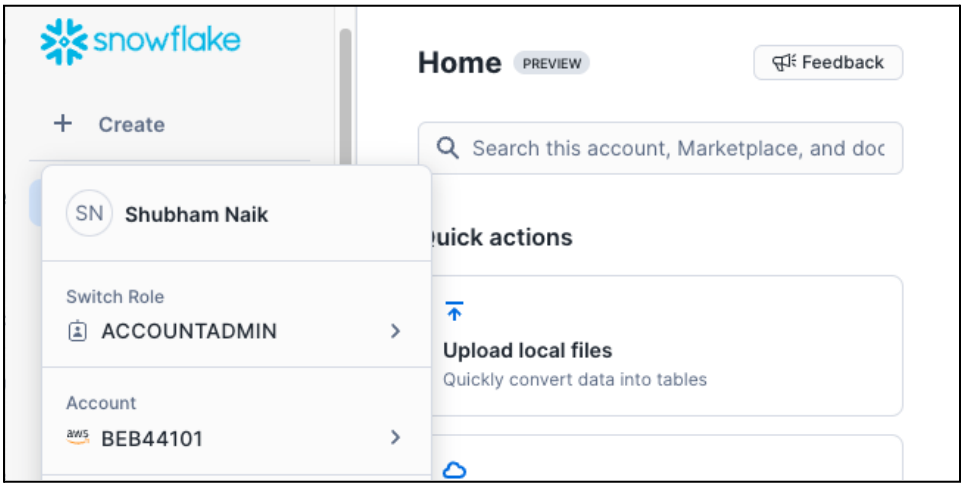
- Install Apache Airflow

```
(airflow_env) (base) shubhamnaik@Shubhams-MacBook-Air ~ % airflow version
/Users/shubhamnaik/airflow_env/lib/python3.12/site-packages/airflow/configuration.py:859 FutureWarnin
g: section/key [core/sql_alchemy_conn] has been deprecated, you should use[database/sql_alchemy_conn]
 instead. Please update your `conf.get*` call to use the new name
2.10.3
```

- Install dbt

```
(base) shubhamnaik@Shubhams-MacBook-Air connection % dbt --version
Core:
  - installed: 1.8.0
  - latest:    1.8.8 - Update available!

  Your version of dbt-core is out of date!
  You can find instructions for upgrading here:
  https://docs.getdbt.com/docs/installation

Plugins:
  - snowflake: 1.8.4 - Up to date!
```

- Set up a Snowflake account

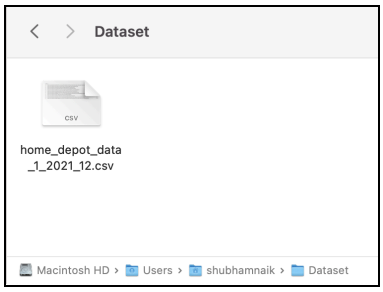


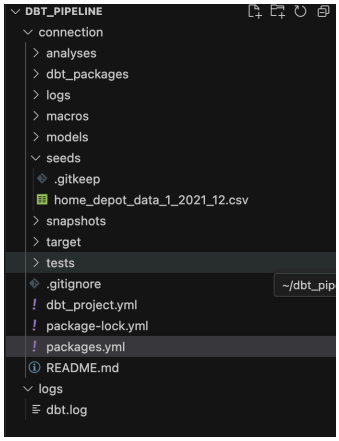
- Install the Snowflake Connector for Python: (pip install snowflake-connector-python)

```
(snowflake_env) (base) shubhamnaik@Shubhams-MacBook-Air ~ % python -c "import snowflake.connector; pr
int(snowflake.connector.__version__)"

3.12.3
```

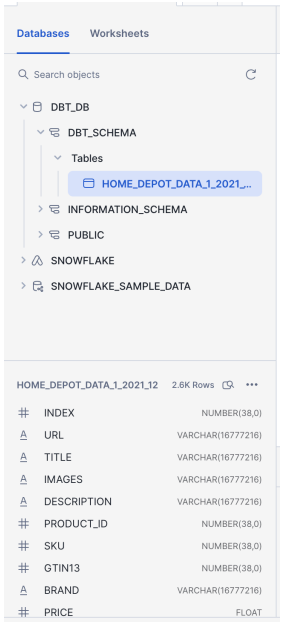
2] Source a dataset:

- Choose an appropriate dataset, such as this one: **Link**
- Store this file in an local filesystem



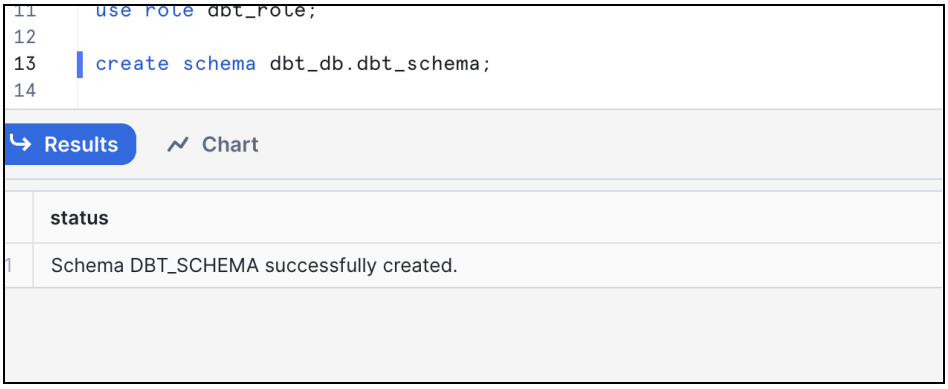



```
(base) shubhamnaik@Shubhams-MacBook-Air connection % dbt seed
01:49:48  Running with dbt=1.8.0
01:49:49  Registered adapter: snowflake=1.8.4
01:49:49  Unable to do partial parsing because saved manifest not found. Starting full parse.
01:49:49  Found 2 models, 1 seed, 4 data tests, 575 macros
01:49:49
01:49:51  Concurrency: 1 threads (target='dev')
01:49:51
01:49:51  1 of 1 START seed file dbt_schema.home_depot_data_1_2021_12 ..... [RUN]
01:49:55  1 of 1 OK loaded seed file dbt_schema.home_depot_data_1_2021_12 ..... [INSERT 2551 in 4.61s]
01:49:55
01:49:55  Finished running 1 seed in 0 hours 0 minutes and 6.08 seconds (6.08s).
01:49:55
01:49:55  Completed successfully
01:49:55
01:49:55  Done. PASS=1 WARN=0 ERROR=0 SKIP=0 TOTAL=1
```

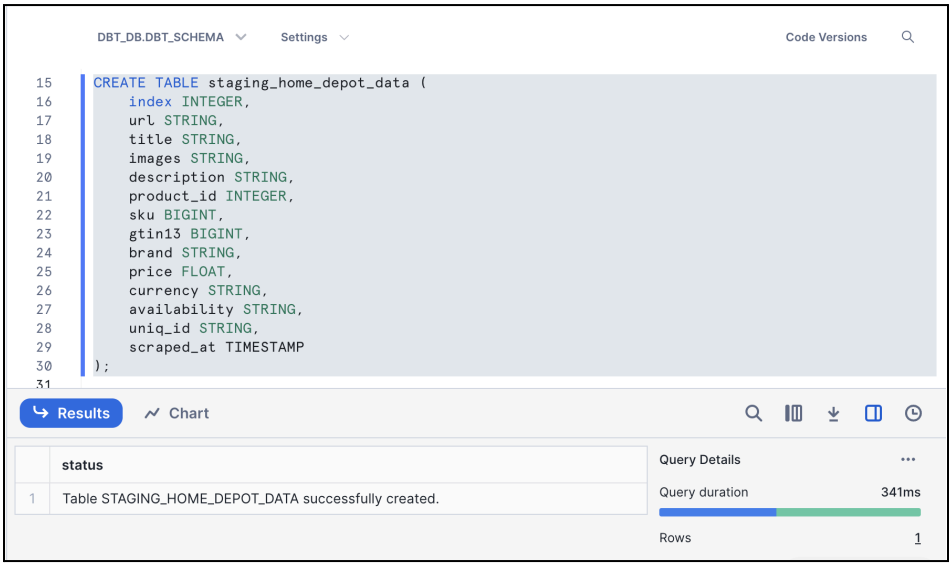


3] Set up Snowflake:

- Create a database and schema for the project



- Create a staging table to receive raw data from the source



4] Implement the Airflow DAG:

- Create a new DAG file in your Airflow `dags` folder

- Define tasks for:

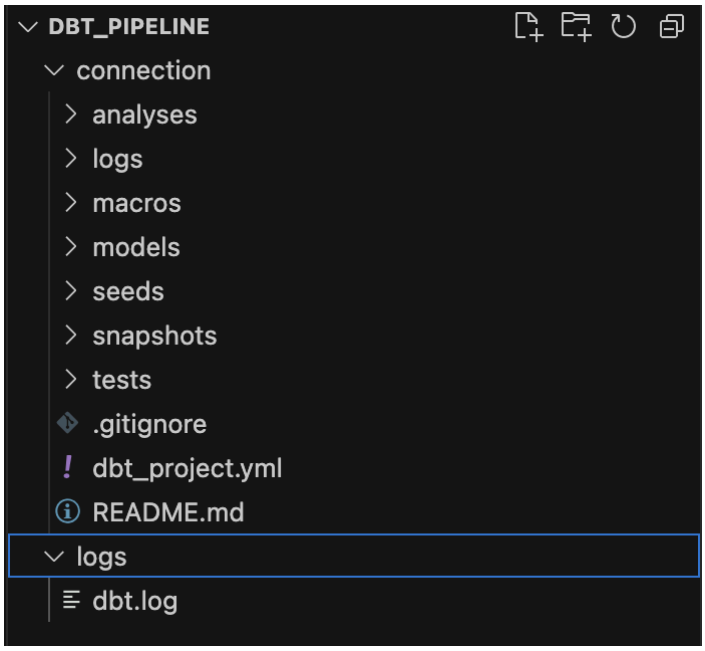
Extracting data from the source

Loading data into Snowflake staging table

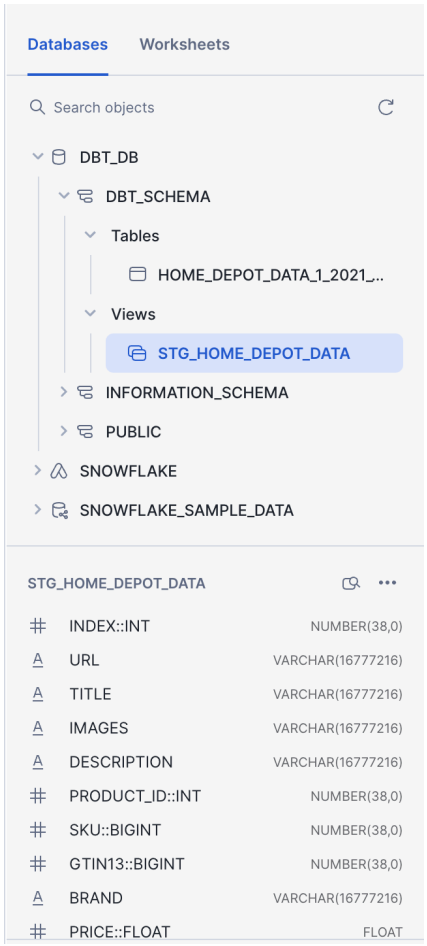
Triggering dbt run and implement error handling

5] Implement dbt models:

- Create a dbt project structure



- Implement a staging model to clean and prepare the data



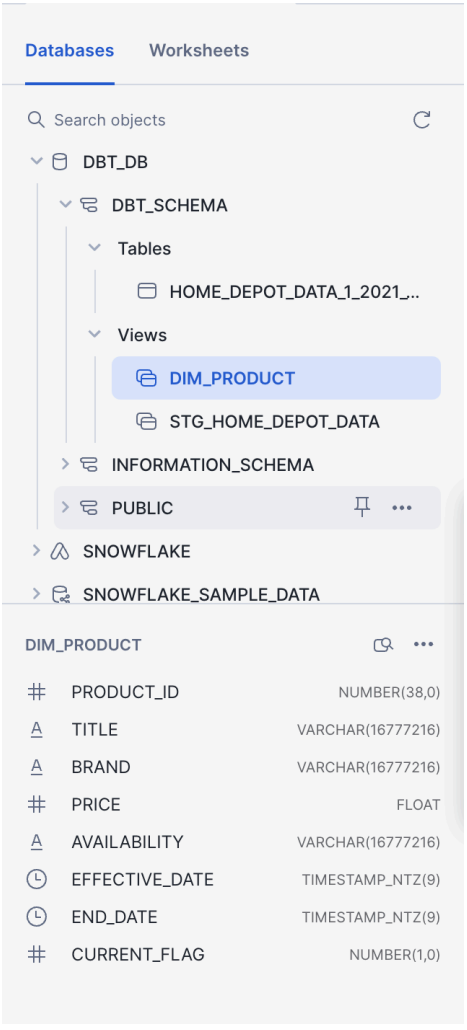
stg_home_depot_data.sql | sources.yml | dim_product.sql

connection > models > staging > stg_home_depot_data.sql

```sql
WITH raw_data AS (
    SELECT
        index::int,
        url,
        title,
        images,
        description,
        product_id::int,
        sku::bigint,
        gtin13::bigint,
        brand,
        price::float,
        currency,
        availability,
        uniq_id,
        scraped_at::timestamp
    FROM {{ source('home_depot_source', 'raw_data') }}
)

SELECT * FROM raw_data
```

```
02:19:53  Running with dbt=1.8.0
02:19:53  Registered adapter: snowflake=1.8.4
02:19:53  [WARNING]: Configuration paths exist in your dbt_project.yml file which do not apply to any resources.
There are 1 unused configuration paths:
- models.connection.example
02:19:53  Found 1 seed, 2 models, 4 data tests, 1 source, 575 macros
02:19:53
02:19:55  Concurrency: 1 threads (target='dev')
02:19:55
02:19:55  1 of 1 START sql view model dbt_schema.stg_home_depot_data ..................... [RUN]
02:19:56  1 of 1 OK created sql view model dbt_schema.stg_home_depot_data ................ [SUCCESS 1 in 1.00s]
02:19:56
02:19:56  Finished running 1 view model in 0 hours 0 minutes and 2.41 seconds (2.41s).
02:19:56
02:19:56  Completed successfully
02:19:56
02:19:56  Done. PASS=1 WARN=0 ERROR=0 SKIP=0 TOTAL=1
(base) shubhamnaik@Shubhams-MacBook-Air connection %
```

- Create a core model that implements SCD Type 2 for the product dimension



stg_home_depot_data.sql | sources.yml | dim_product.sql | dim_product_tests.yml .../core | din

connection > models > core > dim_product.sql

```sql
WITH latest_data AS (
    SELECT
        "PRODUCT_ID",
        "TITLE",
        "BRAND",
        "PRICE",
        "AVAILABILITY",
        "SCRAPED_AT",
        ROW_NUMBER() OVER(PARTITION BY "PRODUCT_ID" ORDER BY "SCRAPED_AT" DESC) AS row_num
    FROM {{ ref('stg_home_depot_data') }}
),

effective_data AS (
    SELECT
        "PRODUCT_ID",
        "TITLE",
        "BRAND",
        "PRICE",
        "AVAILABILITY",
        "SCRAPED_AT" AS effective_date,
        LEAD("SCRAPED_AT") OVER(PARTITION BY "PRODUCT_ID" ORDER BY "SCRAPED_AT") AS end_date,
        CASE WHEN row_num = 1 THEN 1 ELSE 0 END AS current_flag
    FROM latest_data
)

SELECT
    "PRODUCT_ID",
    "TITLE",
    "BRAND",
    "PRICE",
    "AVAILABILITY",
    effective_date,
    end_date,
    current_flag
FROM effective_data
```

```
(base) shubhamnaik@Shubhams-MacBook-Air connection % dbt run --models dim_product

02:57:19  Running with dbt=1.8.0
02:57:19  Registered adapter: snowflake=1.8.4
02:57:19  [WARNING]: Configuration paths exist in your dbt_project.yml file which do not apply to any resources.
There are 1 unused configuration paths:
- models.connection.example
02:57:20  Found 1 seed, 2 models, 4 data tests, 1 source, 575 macros
02:57:20
02:57:21  Concurrency: 1 threads (target='dev')
02:57:21
02:57:21  1 of 1 START sql view model dbt_schema.dim_product ............................. [RUN]
02:57:22  1 of 1 OK created sql view model dbt_schema.dim_product ........................ [SUCCESS 1 in 1.14s]
02:57:22
02:57:22  Finished running 1 view model in 0 hours 0 minutes and 2.86 seconds (2.86s).
02:57:22
02:57:22  Completed successfully
02:57:22
02:57:22  Done. PASS=1 WARN=0 ERROR=0 SKIP=0 TOTAL=1
```

- Write appropriate tests for your models

connection > tests > ! dim_product_tests.yml

```yaml
version: 2
models:
  - name: dim_product
    columns:
      - name: product_id
        tests:
          - not_null
      - name: effective_date
        tests:
          - not_null
      - name: current_flag
        tests:
          - accepted_values:
              values: [0, 1]
          - unique_combination_of_columns:
              - product_id
              - effective_date
```

connection > tests > ! stg_home_depot_data_tests.yml

```yaml
version: 2
models:
  - name: stg_home_depot_data
    columns:
      - name: product_id
        tests:
          - not_null
          - unique
      - name: price
        tests:
          - not_null
```

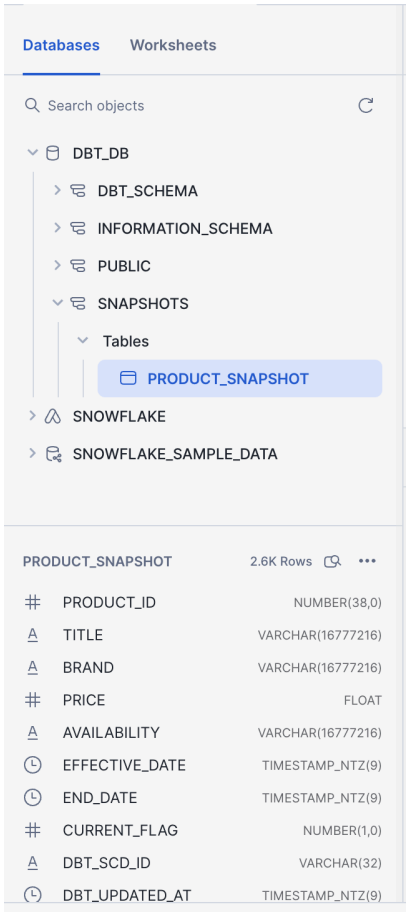
```
(base) shubhamnaik@Shubhams-MacBook-Air connection % dbt test
03:00:15  Running with dbt=1.8.0
03:00:15  Registered adapter: snowflake=1.8.4
03:00:16  [WARNING]: Deprecated functionality
The `tests` config has been renamed to `data_tests`. Please see
https://docs.getdbt.com/docs/build/data-tests#new-data_tests-syntax for more
information.
03:00:16  [WARNING]: Configuration paths exist in your dbt_project.yml file which do not apply to any resources.
There are 1 unused configuration paths:
- models.connection.example
03:00:16  Found 1 seed, 2 models, 4 data tests, 1 source, 575 macros
03:00:16
03:00:17  Concurrency: 1 threads (target='dev')
03:00:17
03:00:17  1 of 4 START test accepted_values_dim_product_current_flag__0__1 ..... [RUN]
03:00:19  1 of 4 PASS accepted_values_dim_product_current_flag__0__1 ..... [PASS in 2.31s]
03:00:19  2 of 4 START test dbt_utils_unique_combination_of_columns_dim_product_product_id__effective_date  [RUN]
03:00:20  2 of 4 PASS dbt_utils_unique_combination_of_columns_dim_product_product_id__effective_date  [PASS in 1.14s]
03:00:20  3 of 4 START test not_null_dim_product_effective_date ..... [RUN]
03:00:21  3 of 4 PASS not_null_dim_product_effective_date ..... [PASS in 1.18s]
03:00:21  4 of 4 START test not_null_dim_product_product_id ..... [RUN]
03:00:22  4 of 4 PASS not_null_dim_product_product_id ..... [PASS in 0.93s]
03:00:22
03:00:22  Finished running 4 data tests in 0 hours 0 minutes and 6.46 seconds (6.46s).
03:00:22
03:00:22  Completed successfully
03:00:22
03:00:22  Done. PASS=4 WARN=0 ERROR=0 SKIP=0 TOTAL=4
```

6] Configure dbt snapshots:

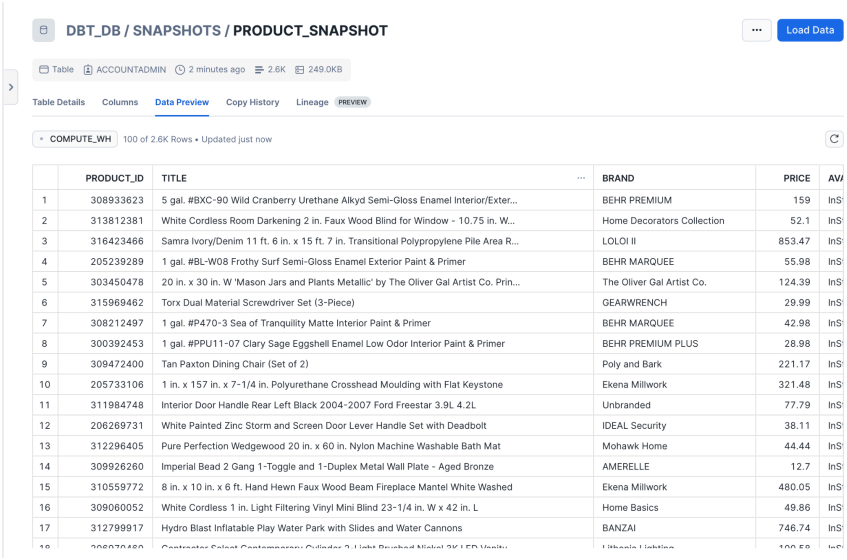
- Implement a snapshot for the product dimension to track historical changes

connection > snapshots > product_snapshot.sql

```sql
{% snapshot product_snapshot %}

SELECT
    PRODUCT_ID,
    TITLE,
    BRAND,
    PRICE,
    AVAILABILITY,
    EFFECTIVE_DATE,
    END_DATE,
    CURRENT_FLAG
FROM {{ ref('dim_product') }}

{% endsnapshot %}

```




```
(base) shubhamnaik@Shubhams-MacBook-Air connection % dbt snapshot

03:31:16  Running with dbt=1.8.0
03:31:16  Registered adapter: snowflake=1.8.4
03:31:16  Unable to do partial parsing because a project config has changed
03:31:17  [WARNING]: Deprecated functionality
The `tests` config has been renamed to `data_tests`. Please see
https://docs.getdbt.com/docs/build/data-tests#new-data_tests-syntax for more
information.
03:31:17  [WARNING]: Configuration paths exist in your dbt_project.yml file which do not apply to any resources.
There are 1 unused configuration paths:
- models.connection.example
03:31:17  Found 2 models, 1 snapshot, 1 seed, 4 data tests, 1 source, 575 macros
03:31:17
03:31:19  Concurrency: 1 threads (target='dev')
03:31:19
03:31:19  1 of 1 START snapshot DBT_DB.snapshots.product_snapshot ..... [RUN]
03:31:22  1 of 1 OK snapshotted DBT_DB.snapshots.product_snapshot ..... [SUCCESS 1 in 2.68s]
03:31:22
03:31:22  Finished running 1 snapshot in 0 hours 0 minutes and 4.81 seconds (4.81s).
03:31:22
03:31:22  Completed successfully
03:31:22
03:31:22  Done. PASS=1 WARN=0 ERROR=0 SKIP=0 TOTAL=1
```



	PRODUCT_ID	TITLE	BRAND	PRICE	AV
1	308933623	5 gal. #BXC-90 Wild Cranberry Urethane Alkyd Semi-Gloss Enamel Interior/Exter...	BEHR PREMIUM	159	In
2	313812381	White Cordless Room Darkening 2 in. Faux Wood Blind for Window - 10.75 in. W...	Home Decorators Collection	52.1	In
3	316423466	Samra Ivory/Denim 11 ft. 6 in. x 15 ft. 7 in. Transitional Polypropylene Pile Area R...	LOLOI II	853.47	In
4	205239289	1 gal. #BL-W08 Frothy Surf Semi-Gloss Enamel Exterior Paint & Primer	BEHR MARQUEE	55.98	In
5	303450478	20 in. x 30 in. W 'Mason Jars and Plants Metallic' by The Oliver Gal Artist Co. Prin...	The Oliver Gal Artist Co.	124.39	In
6	315969462	Torx Dual Material Screwdriver Set (3-Piece)	GEARWRENCH	29.99	In
7	308212497	1 gal. #P470-3 Sea of Tranquility Matte Interior Paint & Primer	BEHR MARQUEE	42.98	In
8	300392453	1 gal. #PPU11-07 Clary Sage Eggshell Enamel Low Odor Interior Paint & Primer	BEHR PREMIUM PLUS	28.98	In
9	309472400	Tan Paxton Dining Chair (Set of 2)	Poly and Bark	221.17	In
10	205733106	1 in. x 157 in. x 7-1/4 in. Polyurethane Crosshead Moulding with Flat Keystone	Ekena Millwork	321.48	In
11	311984748	Interior Door Handle Rear Left Black 2004-2007 Ford Freestar 3.9L 4.2L	Unbranded	77.79	In
12	206269731	White Painted Zinc Storm and Screen Door Lever Handle Set with Deadbolt	IDEAL Security	38.11	In
13	312296405	Pure Perfection Wedgewood 20 in. x 60 in. Nylon Machine Washable Bath Mat	Mohawk Home	44.44	In
14	309926260	Imperial Bead 2 Gang 1-Toggle and 1-Duplex Metal Wall Plate - Aged Bronze	AMERELLE	12.7	In
15	310559772	8 in. x 10 in. x 6 ft. Hand Hewn Faux Wood Beam Fireplace Mantel White Washed	Ekena Millwork	480.05	In
16	309060052	White Cordless 1 in. Light Filtering Vinyl Mini Blind 23-1/4 in. W x 42 in. L	Home Basics	49.86	In
17	312799917	Hydro Blast Inflatable Play Water Park with Slides and Water Cannons	BANZAI	746.74	In

7] Integrate dbt with Airflow:

- Use the `dbt` operator in Airflow to run your dbt models

Dockerfile | requirements.txt | dag_dbt.py 2 | packages.yml

dags > dag_dbt.py > ...

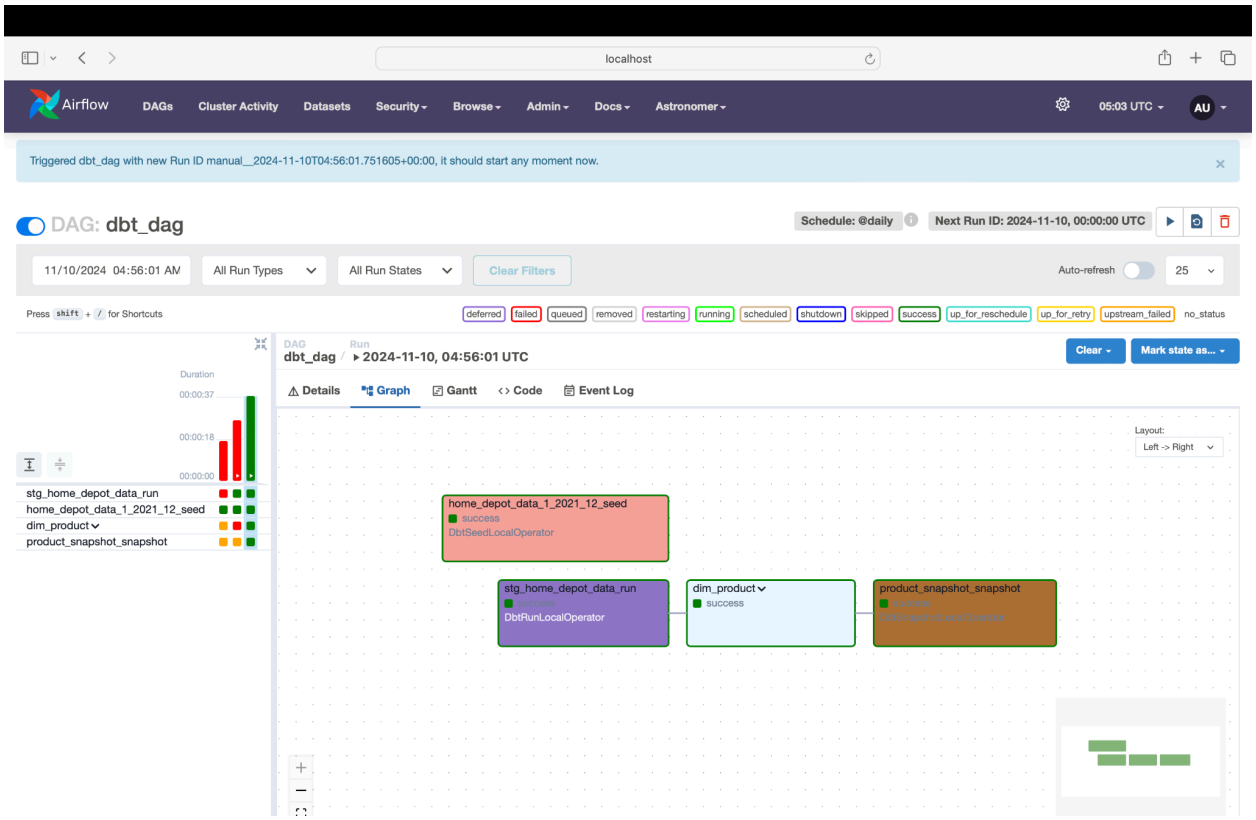
```python
import os
from datetime import datetime

from cosmos import DbtDag, ProjectConfig, ProfileConfig, ExecutionConfig
from cosmos.profiles import SnowflakeUserPasswordProfileMapping


profile_config = ProfileConfig(
    profile_name="default",
    target_name="dev",
    profile_mapping=SnowflakeUserPasswordProfileMapping(
        conn_id="snowflake_conn",
        profile_args={"database": "dbt_db", "schema": "dbt_schema"},
    )
)

dbt_snowflake_dag = DbtDag(
    project_config=ProjectConfig("/usr/local/airflow/dags/dbt/connection",),
    operator_args={"install_deps": True},
    profile_config=profile_config,
    execution_config=ExecutionConfig(dbt_executable_path=f"{os.environ['AIRFLOW_HOME']}/dbt_venv/bin/dbt",),
    schedule_interval="@daily",
    start_date=datetime(2023, 9, 10),
    catchup=False,
    dag_id="dbt_dag",
)
```

8] Test the entire pipeline:

- Run the Airflow DAG

- Verify data in Snowflake

- Make changes to the source data and re-run the pipeline to test SCD implementation



DBT_DB / DBT_SCHEMA / DIM_PRODUCT

View · ACCOUNTADMIN · 9 minutes ago

View Details | Columns | Data Preview | Lineage PREVIEW

COMPUTE_WH · Updated just now

	PRODUCT_ID	TITLE	BRAND	PRICE	AVA
1	308933623	5 gal. #BXC-90 Wild Cranberry Urethane Alkyd Semi-Gloss Enamel Interior/Exter...	BEHR PREMIUM	159	InS
2	313812381	White Cordless Room Darkening 2 in. Faux Wood Blind for Window - 10.75 in. W...	Home Decorators Collection	52.1	InS
3	316423466	Samra Ivory/Denim 11 ft. 6 in. x 15 ft. 7 in. Transitional Polypropylene Pile Area R...	LOLOI II	853.47	InS
4	205239289	1 gal. #BL-W08 Frothy Surf Semi-Gloss Enamel Exterior Paint & Primer	BEHR MARQUEE	55.98	InS
5	303450478	20 in. x 30 in. W 'Mason Jars and Plants Metallic' by The Oliver Gal Artist Co. Prin...	The Oliver Gal Artist Co.	124.39	InS
6	315969462	Torx Dual Material Screwdriver Set (3-Piece)	GEARWRENCH	29.99	InS
7	308212497	1 gal. #P470-3 Sea of Tranquility Matte Interior Paint & Primer	BEHR MARQUEE	42.98	InS
8	300392453	1 gal. #PPU11-07 Clary Sage Eggshell Enamel Low Odor Interior Paint & Primer	BEHR PREMIUM PLUS	28.98	InS
9	309472400	Tan Paxton Dining Chair (Set of 2)	Poly and Bark	221.17	InS
10	205733106	1 in. x 157 in. x 7-1/4 in. Polyurethane Crosshead Moulding with Flat Keystone	Ekena Millwork	321.48	InS
11	311984748	Interior Door Handle Rear Left Black 2004-2007 Ford Freestar 3.9L 4.2L	Unbranded	77.79	InS
12	206269731	White Painted Zinc Storm and Screen Door Lever Handle Set with Deadbolt	IDEAL Security	38.11	InS
13	312296405	Pure Perfection Wedgewood 20 in. x 60 in. Nylon Machine Washable Bath Mat	Mohawk Home	44.44	InS
14	309926260	Imperial Bead 2 Gang 1-Toggle and 1-Duplex Metal Wall Plate - Aged Bronze	AMERELLE	12.7	InS
15	310559772	8 in. x 10 in. x 6 ft. Hand Hewn Faux Wood Beam Fireplace Mantel White Washed	Ekena Millwork	480.05	InS
16	309060052	White Cordless 1 in. Light Filtering Vinyl Mini Blind 23-1/4 in. W x 42 in. L	Home Basics	49.86	InS
17	312799917	Hydro Blast Inflatable Play Water Park with Slides and Water Cannons	BANZAI	746.74	InS
18	206970460	Contractor Select Contemporary Cylinder 2-Light Brushed Nickel 3K LED Vanity	Lithonia Lighting	100.58	InS

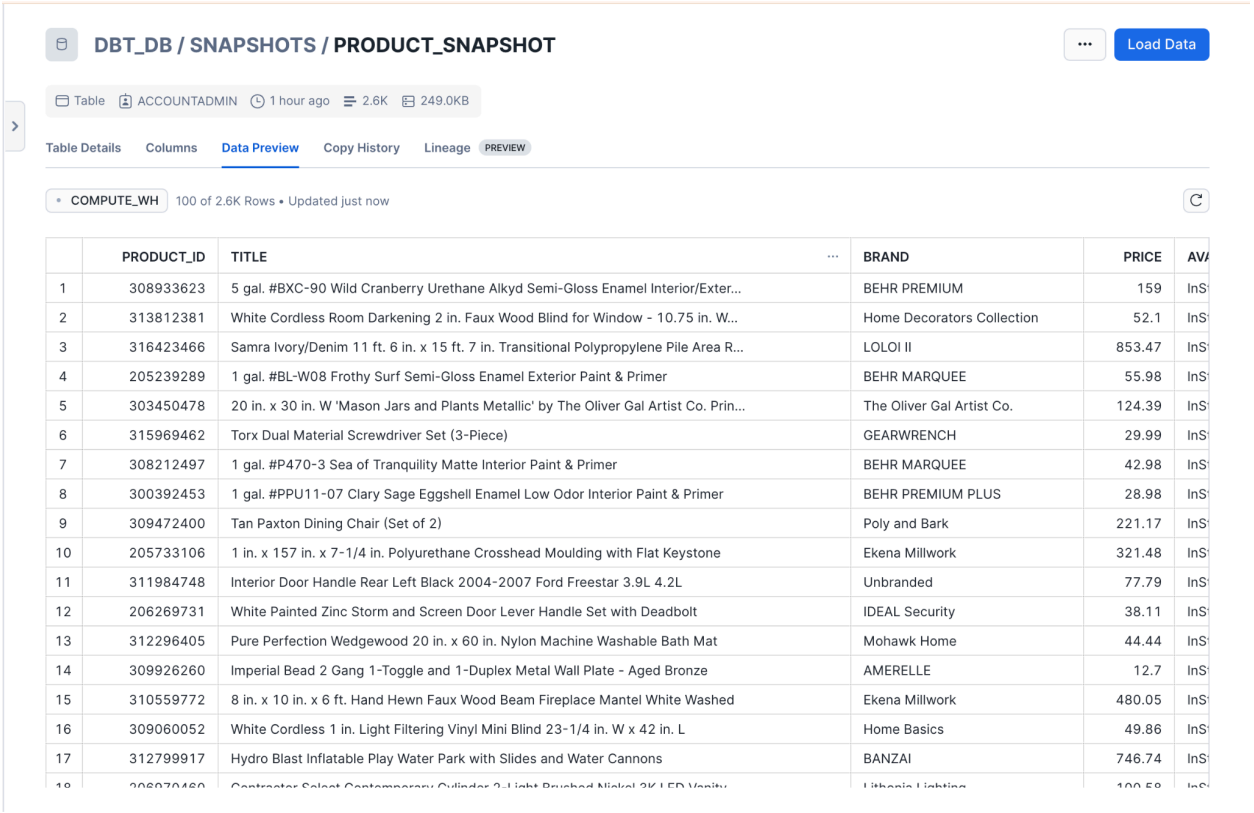
DBT_DB / DBT_SCHEMA / STG_HOME_DEPOT_DATA

View · ACCOUNTADMIN · 10 minutes ago

View Details | Columns | Data Preview | Lineage PREVIEW

COMPUTE_WH · Updated just now

	INDEX	URL	TITLE
1	0	https://www.homedepot.com/p/Carhartt-Men-s-3X-Large-Carbon-Heather-Co...	Men's 3X Large Carbon Heather Cotton/Polyester Rain Defender F
2	1	https://www.homedepot.com/p/Turmode-30-ft-RP-TNC-Female-to-RP-TNC-Ma...	Turmode 30 ft. RP TNC Female to RP TNC Male Adapter Cable
3	2	https://www.homedepot.com/p/Carolina-Pet-Company-Large-Tapestry-Bolster...	Large Tapestry Bolster Bed
4	3	https://www.homedepot.com/p/16-Gauge-Sinks-Vessel-Sink-in-White-with-Fa...	16-Gauge-Sinks Vessel Sink in White with Faucet
5	4	https://www.homedepot.com/p/Adtec-Men-s-Crazy-Horse-9-Logger-Boot-Ste...	Men's Crazy Horse 9" Logger Boot - Steel Toe - Black Size 10.5(W
6	5	https://www.homedepot.com/p/HomeRoots-Mariana-6-ft-Multi-Color-3-Panel-...	Mariana 6 ft. Multi-Color 3-Panel Screen Divider
7	6	https://www.homedepot.com/p/BEHR-PRO-5-gal-650C-2-Powdery-Mist-Semi-...	5 gal. #650C-2 Powdery Mist Semi-Gloss Interior Paint
8	7	https://www.homedepot.com/p/DEWALT-7-8-in-x-4-1-2-in-x-0-045-in-Metal-...	7/8 in. x 4-1/2 in. x 0.045 in. Metal and Stainless Cutting Wheel (5
9	8	https://www.homedepot.com/p/Titan-Lighting-Ring-Gold-Bar-Cart-TN-892747...	Ring Gold Bar Cart
10	9	https://www.homedepot.com/p/Benjara-Traditional-Silver-Wooden-Vanity-Tabl...	Traditional Silver Wooden Vanity Table
11	10	https://www.homedepot.com/p/Ply-Gem-15-in-x-59-in-Open-Louvered-Polypr...	15 in. x 59 in. Open Louvered Polypropylene Shutters Pair in Pepp
12	11	https://www.homedepot.com/p/BEHR-PREMIUM-PLUS-1-qt-350F-7-Wild-Mus...	1 qt. #350F-7 Wild Mushroom Semi-Gloss Enamel Low Odor Inter
13	12	https://www.homedepot.com/p/Anthracite-Cordless-Light-Filtering-Fabric-Cell...	Anthracite Cordless Light Filtering Fabric Cellular Shade 9/16 in. S
14	13	https://www.homedepot.com/p/Luverne-SlimGrip-78-Inch-Black-Aluminum-Tru...	SlimGrip 78-Inch Black Aluminum Truck Running Boards, Select Ra
15	14	https://www.homedepot.com/p/Ekena-Millwork-6-in-x-28-in-x-28-in-Douglas-...	6 in. x 28 in. x 28 in. Douglas Fir Balboa Arts and Crafts Rough Sa
16	15	https://www.homedepot.com/p/Home-Decorators-Collection-Espresso-Cordles...	Espresso Cordless Room Darkening 2.5 in. Premium Faux Wood Bl
17	16	https://www.homedepot.com/p/BEHR-PREMIUM-PLUS-5-gal-BL-W10-Maui-Mi...	5 gal. #BL-W10 Maui Mist Semi-Gloss Enamel Low Odor Interior P
18	17	https://www.homedepot.com/p/WAREHOUSE-OF-TIFFANY-Floyn-26-in-Indoor...	Floyn 26 in. Indoor Chrome Remote Controlled Fandelier with Ligh



	PRODUCT_ID	TITLE	BRAND	PRICE	AVA
1	308933623	5 gal. #BXC-90 Wild Cranberry Urethane Alkyd Semi-Gloss Enamel Interior/Exter...	BEHR PREMIUM	159	InS
2	313812381	White Cordless Room Darkening 2 in. Faux Wood Blind for Window - 10.75 in. W...	Home Decorators Collection	52.1	InS
3	316423466	Samra Ivory/Denim 11 ft. 6 in. x 15 ft. 7 in. Transitional Polypropylene Pile Area R...	LOLOI II	853.47	InS
4	205239289	1 gal. #BL-W08 Frothy Surf Semi-Gloss Enamel Exterior Paint & Primer	BEHR MARQUEE	55.98	InS
5	303450478	20 in. x 30 in. W 'Mason Jars and Plants Metallic' by The Oliver Gal Artist Co. Prin...	The Oliver Gal Artist Co.	124.39	InS
6	315969462	Torx Dual Material Screwdriver Set (3-Piece)	GEARWRENCH	29.99	InS
7	308212497	1 gal. #P470-3 Sea of Tranquility Matte Interior Paint & Primer	BEHR MARQUEE	42.98	InS
8	300392453	1 gal. #PPU11-07 Clary Sage Eggshell Enamel Low Odor Interior Paint & Primer	BEHR PREMIUM PLUS	28.98	InS
9	309472400	Tan Paxton Dining Chair (Set of 2)	Poly and Bark	221.17	InS
10	205733106	1 in. x 157 in. x 7-1/4 in. Polyurethane Crosshead Moulding with Flat Keystone	Ekena Millwork	321.48	InS
11	311984748	Interior Door Handle Rear Left Black 2004-2007 Ford Freestar 3.9L 4.2L	Unbranded	77.79	InS
12	206269731	White Painted Zinc Storm and Screen Door Lever Handle Set with Deadbolt	IDEAL Security	38.11	InS
13	312296405	Pure Perfection Wedgewood 20 in. x 60 in. Nylon Machine Washable Bath Mat	Mohawk Home	44.44	InS
14	309926260	Imperial Bead 2 Gang 1-Toggle and 1-Duplex Metal Wall Plate - Aged Bronze	AMERELLE	12.7	InS
15	310559772	8 in. x 10 in. x 6 ft. Hand Hewn Faux Wood Beam Fireplace Mantel White Washed	Ekena Millwork	480.05	InS
16	309060052	White Cordless 1 in. Light Filtering Vinyl Mini Blind 23-1/4 in. W x 42 in. L	Home Basics	49.86	InS
17	312799917	Hydro Blast Inflatable Play Water Park with Slides and Water Cannons	BANZAI	746.74	InS
18	206970460	Contractor Select Contemporary Cylinder 2-Light Brushed Nickel 3K LED Vanity	Lithonia Lighting	100.58	InS

Leverage Snowflake Zero-Copy Cloning

9.1: [1 point] Development Environment Cloning

Objective: Create a development environment using zero-copy cloning to test changes without affecting the production data.

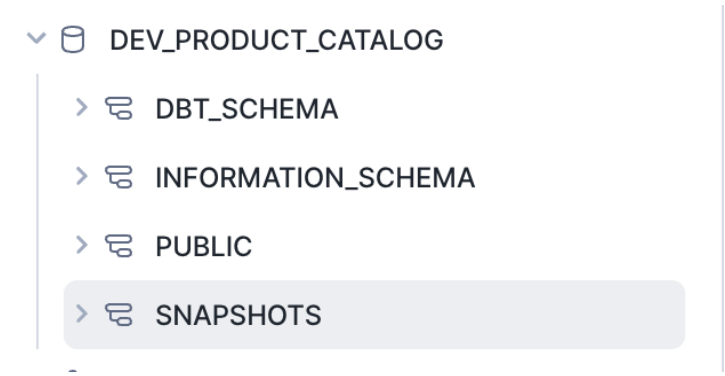
Steps:

1. Clone the production database to create a development environment:

```
47
48  CREATE DATABASE dev_product_catalog CLONE dbt_db;
49
50
51
```

Results | Chart

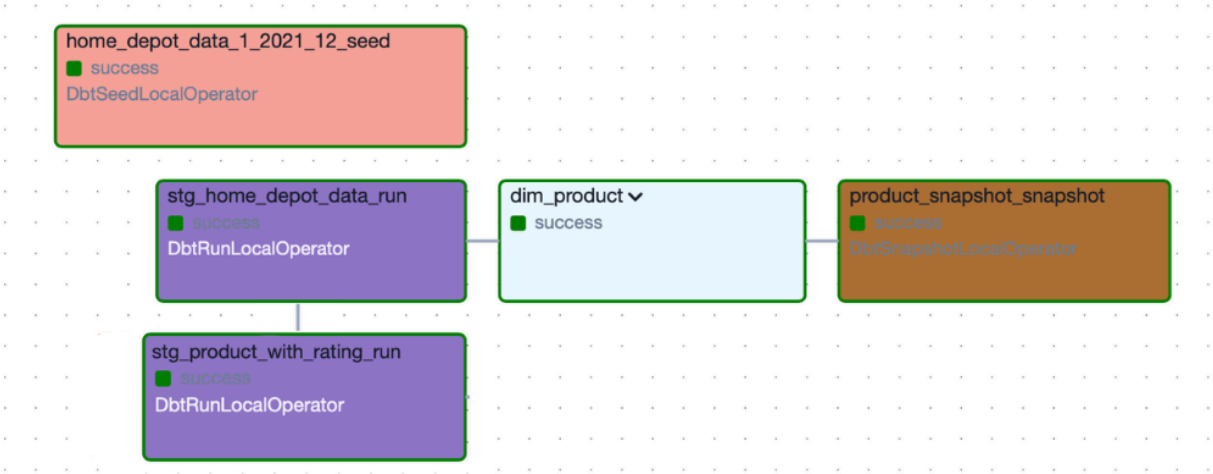
status
Database DEV_PRODUCT_CATALOG successfully created.



1. Implement a dbt model in the development environment to test a new feature (e.g., adding a product rating column).
2. Create a new Airflow DAG to run dbt models in the development environment.
3. After successful testing, implement the changes in the production environment.

dags > dbt > connection > models > staging > stg_product_with_rating.sql

```sql
-- models/staging/stg_product_with_rating.sql
SELECT
    product_id,
    title,
    description,
    price,
    CASE
        WHEN review_score IS NOT NULL THEN review_score / 2  -- Sample transformation for rating
        ELSE 0
    END AS product_rating  -- New column for rating
FROM
    {{ source('product_catalog', 'source_product_table') }}
```



9.2: Point-in-Time Analysis

Objective: Use zero-copy cloning to create a snapshot of the database for historical analysis.

Steps:

Create a clone of the product catalog database at a specific point in time:

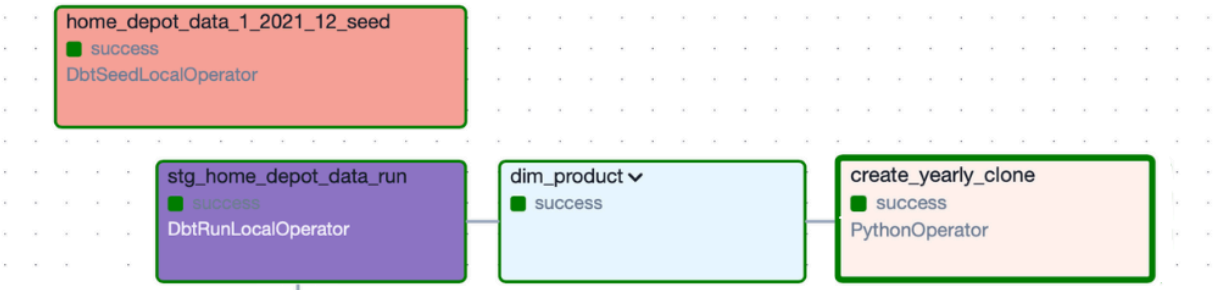
1. Implement a dbt model that performs year-over-year analysis using this cloned database.
2. Create an Airflow task to generate this clone at the end of each year:

```
51
52  CREATE DATABASE product_catalog_snapshot_2024 CLONE dbt_db;
53
54
```

Results | Chart

	status
1	Database PRODUCT_CATALOG_SNAPSHOT_2024 successfully created.

```python
with DAG(
    'yearly_clone_dag',
    schedule_interval='@yearly',
    start_date=datetime(2023, 12, 31),
    catchup=False
) as dag:

    clone_task = PythonOperator(
        task_id='create_yearly_clone',
        python_callable=create_yearly_clone
    )
```



9.3: Rapid Prototyping with Cloning

Objective: Use zero-copy cloning to quickly prototype and test new data models. Steps:

Clone a specific schema for prototyping:

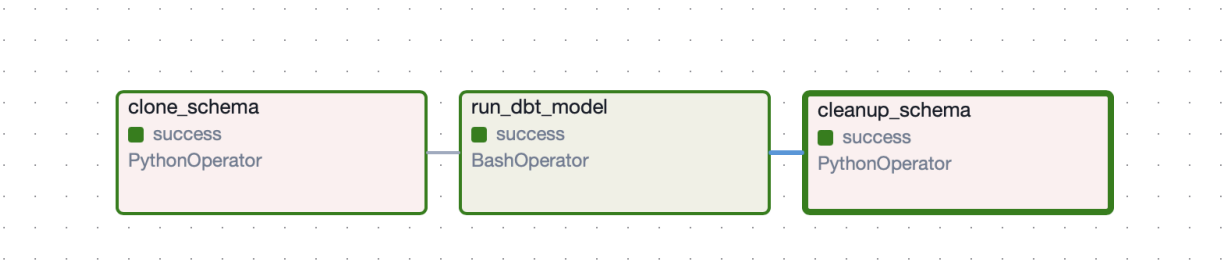
1. Implement a new dbt model in this cloned schema to test a complex transformation.
2. Create an Airflow task to clone the schema, run the prototype model, and clean up

```
53
54  CREATE SCHEMA dbt_prototype_schema CLONE dbt_db.dbt_schema;
55
56
57
58
```

Results | Chart

	status
1	Schema DBT_PROTOTYPE_SCHEMA successfully created.

```python
# Task to clone the schema
clone_task = PythonOperator(
    task_id='clone_schema',
    python_callable=clone_schema
)

# Task to run the dbt model in the prototype schema
dbt_run_task = BashOperator(
    task_id='run_dbt_model',
    bash_command='dbt run --models prototyping.prototyped_complex_transformation --target dev'
)

# Task to clean up the cloned schema after testing
cleanup_task = PythonOperator(
    task_id='cleanup_schema',
    python_callable=cleanup_schema
)

# Define the task dependencies
clone_task >> dbt_run_task >> cleanup_task
```



9.4: Backup and Restore Strategy

Objective: Implement a backup and restore strategy using zero-copy cloning.

Steps:

- Create a daily backup of the production database:

- Implement an Airflow task to manage backups, keeping only the last 7 days:
- Implement a restore procedure using the most recent backup.

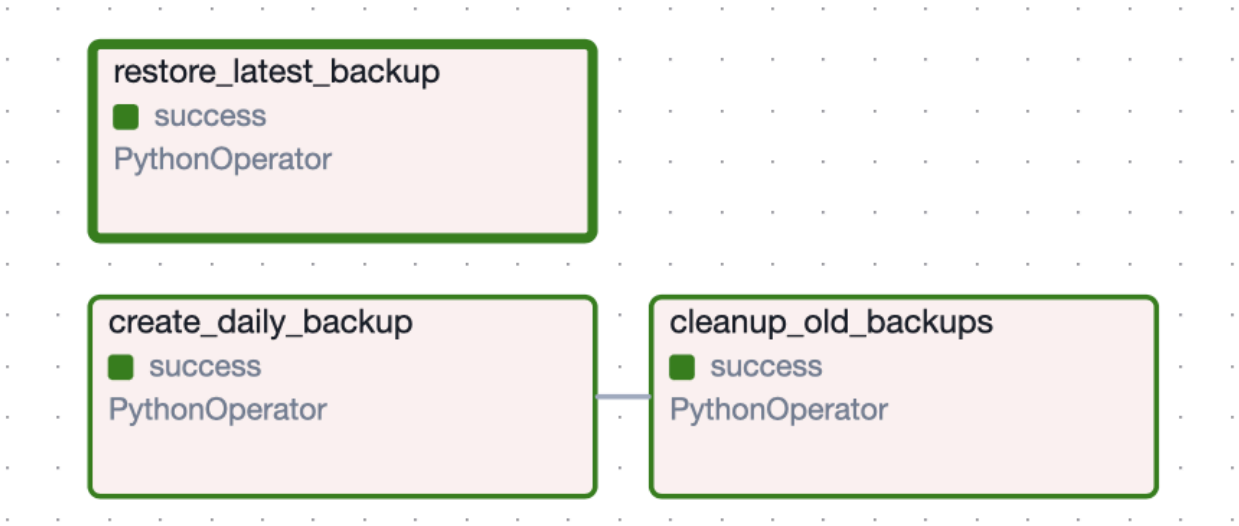
```
CREATE DATABASE product_catalog_backup_2024_11_09 CLONE dbt_db;
```

Results | Chart

status
Database PRODUCT_CATALOG_BACKUP_2024_11_09 successfully created.

```python
def restore_latest_backup():
    conn = snowflake.connector.connect(
        user='<user>',
        password='<password>',
        account='<account>',
        warehouse='<warehouse>'
    )
    cursor = conn.cursor()
    # Get the most recent backup
    cursor.execute("SHOW DATABASES LIKE 'product_catalog_backup_%';")
    databases = cursor.fetchall()
    latest_backup = max(databases, key=lambda db: datetime.strptime(db[1].split('_')[-3:], '%Y_%m_%d'))
    latest_backup_name = latest_backup[1]

    # Creat
    cursor.                                  (function) close: Any
    E product_catalog_restored CLONE {latest_backup_name};")
    cursor.close()
    conn.close()

restore_task = PythonOperator(
    task_id='restore_latest_backup',
    python_callable=restore_latest_backup,
    dag=dag  # Adding it to the existing DAG if needed
)
```

```
    # Filter backups older than 7 days
    retention_date = datetime.now() - timedelta(days=7)
    for db in databases:
        backup_name = db[1]  # Assuming database name is the second item
        date_str = backup_name.split('_')[-3:]  # Extract the date part
        backup_date = datetime.strptime('_'.join(date_str), '%Y_%m_%d')
        if backup_date < retention_date:
            cursor.execute(f"DROP DATABASE IF EXISTS {backup_name};")

    cursor.close()
    conn.close()

with DAG(
    'daily_backup_management_dag',
    schedule_interval='@daily',
    start_date=datetime(2023, 1, 1),
    catchup=False
) as dag:

    # Task to create a daily backup
    backup_task = PythonOperator(
        task_id='create_daily_backup',
        python_callable=create_daily_backup
    )

    # Task to clean up backups older than 7 days
    cleanup_task = PythonOperator(
        task_id='cleanup_old_backups',
        python_callable=cleanup_old_backups
    )


    # Define task dependencies
    backup_task >> cleanup_task
```



10. Experiment with various options and describe 5 key takeaways.

1. Data Ingestion Flexibility with Airflow and dbt Seed

Loading CSVs locally with dbt seed simplifies the ingestion of data. It puts the loading of data in one location within the dbt project. Then, if you add in orchestration with Airflow, these static datasets are managed with ease during development without relying on cloud storage.

2. Managing Historical Data with SCD Type 2 in dbt

SCD Type 2 is best suited for use cases where historic data needs to be kept. It lets you track changes over a specific time. The modular approach of dbt helped in keeping the models, which needed to track historical data, isolated from the main pipeline and made it easier to manage and read.

3. Snapshotting and Point-in-Time Analysis

Snowflake's zero-copy cloning means that you can very efficiently snapshot at any moment in time, not doubling storage costs, and can be fitted for time-based data analysis. The process will also enable you to perform such temporal comparisons as year-over-year growth, which informs a lot about the product trends over time.

4. Rapid Prototyping with Schema Cloning

Zero-copy cloning is super helpful in quick prototyping and testing. It gives developers the chance to experiment in environments that resemble production; this speeds up development without breaking the stability in the main database.

5. Automated Backup and Restore Strategy

The scheduled retention policies included in an automated backup strategy greatly reduce the need to intervene manually and minimize the risk of losing data. I did perform effective and efficient backups, both time- and space-efficient, using zero-copy cloning from Snowflake.

11. **If given a chance, how will you extend this exercise to make the learning experience better?**

Integrate Data Quality Checks with Automated Alerts

Include dbt tests for data quality checks - e.g., null values, valid ranges - and add airflow alerts in case of test failures. The inclusion here emphasizes that data quality is critical in production pipelines, while setting up automated monitoring is considered essential in maintaining reliable data pipelines.

Deploy the Pipeline on a Cloud Platform

Guide students to deploy the pipeline to a cloud platform, such as AWS or GCP, using cloud storage for dbt seeds and Snowflake as the data warehouse. By working on cloud-based deployment, the student learns about the management of cloud resources, security considerations, and scalability. This bridges a gap between local development and the production environment that the real world faces.

Implement Incremental Loading to Optimize Performance

Introduce incremental loading in dbt that will update only the changed data to minimize the volumes being processed day in and day out. This exercise shows students how to optimize pipeline performance by handling larger datasets more efficiently. This is an important skill in being able to handle large volumes of data when working on applications in the real world.