NYC Green Taxi Analytics Lab

– Follow-Along Instructions

# 🔽 Download and Prepare Seed Files

1. Download the following files to your computer from [here](https://stratanconsulting-my.sharepoint.com/:f:/p/asim/EmUDundRX2JAiMQ_h7R0EE0BY1Ka64DNscSxz5LrV68t7Q?e=NrRa5v):  
   - green\_tripdata\_2021\_07.csv  
   - taxi\_zone\_lookup.csv
2. Open dbt Studio and go to your project.
3. On the left-hand side click on Studio
4. In File Explorer >> Under the seed/ folder, create files with the exact names:  
   - green\_tripdata\_2021\_07.csv  
   - taxi\_zone\_lookup.csv
5. Open the two downloaded csvs in notepad
6. Copy and paste the contents of each downloaded CSV into their newly created respective files in dbt.

# ⚙️ Configure Seeds

1. In file explorer find the dbt\_project.yml file.

2. In your dbt\_project.yml file, add the following configuration at the very end

seeds:  
 my\_new\_project:  
 +schema: raw

Note: my\_new\_project should match the name at the top of your YAML file.

3. Run dbt seed from the command line:

dbt seed

# 🔍 Validate Seed Load

* Open NeonDB Console
* Click on your project
* Navigate to SQL Editor
* Click on + sign for new query window
* Run this SQL to validate that the tables were loaded (2 rows expected):

SELECT table\_schema, table\_name  
FROM information\_schema.tables  
WHERE table\_name LIKE '%tripdata%' OR table\_name LIKE '%lookup%';

# 🔍 Explore the Raw Data in Neon

* In the SQL Editor run the following queries to preview the data one by one. Replace schema name dbt\_smubashir\_raw with the schema name that your environment provided to you.

-- Total count should be 83691  
SELECT COUNT(\*) FROM dbt\_smubashir\_raw.green\_tripdata\_2021\_07;  
  
-- Preview the lookup data  
SELECT \* FROM dbt\_smubashir\_raw.taxi\_zone\_lookup LIMIT 10;  
  
-- Explore bogus rows  
SELECT \* FROM dbt\_smubashir\_raw.green\_tripdata\_2021\_07   
WHERE total\_amount = 0 AND trip\_distance = 0;  
  
-- Explore usable rows  
SELECT COUNT(\*) FROM dbt\_smubashir\_raw.green\_tripdata\_2021\_07   
WHERE total\_amount > 0 AND trip\_distance > 0;  
-- Total count should be 80034

# 🛠 Create the Staging Layer (Transform)

1. Navigate to DBT >> Studio
2. In the file explorer, create a folder named staging inside models/
3. Create file: models/staging/stg\_green\_tripdata.sql
4. Copy the following code into your newly created file.

-- models/staging/raw/stg\_green\_tripdata.sql  
{{ config(schema='stg', materialized='view') }}  
  
SELECT  
 CAST("VendorID" AS INT) AS vendor\_id,  
 "lpep\_pickup\_datetime",  
 "lpep\_dropoff\_datetime",  
 "store\_and\_fwd\_flag",  
 "RatecodeID",  
 "PULocationID",  
 "DOLocationID",  
 "passenger\_count",  
 "trip\_distance",  
 "fare\_amount",  
 "extra",  
 "mta\_tax",  
 "tip\_amount",  
 "tolls\_amount",  
 "ehail\_fee",  
 "improvement\_surcharge",  
 "total\_amount",  
 "payment\_type",  
 "trip\_type",  
 "congestion\_surcharge"  
FROM {{ source('raw', 'green\_tripdata\_2021\_07') }}  
WHERE "total\_amount" > 0  
 AND "trip\_distance" > 0

1. Create another file: models/staging/stg\_zone\_lookup.sql
2. Copy the following code into new file

-- models/staging/raw/stg\_zone\_lookup.sql  
{{ config(schema='stg', materialized='view') }}  
  
SELECT  
 "LocationID",  
 "Borough",  
 "Zone",  
 "service\_zone"  
FROM {{ source('raw', 'taxi\_zone\_lookup') }}

1. Create file: models/staging/src.yml
2. Copy the following code into the new file

version: 2  
  
sources:  
 - name: raw  
 schema: dbt\_smubashir\_raw  
 tables:  
 - name: green\_tripdata\_2021\_07  
 - name: taxi\_zone\_lookup

# 🚀 Run the Transformation

* In the CLI at the bottom of the screen run the following

dbt run

1. Go to NeonDB console and confirm schema `dbt\_<your\_schema>\_stg` and views created.
2. Run the following queries to confirm and explore data: Change the schema name to yours. The count should be 80034, the same as previous query in the raw layer.

SELECT COUNT(\*) FROM dbt\_smubashir\_stg.stg\_green\_tripdata;

***SELECT \* FROM dbt\_smubashir\_stg.stg\_green\_tripdata limit 1000;***

# 📦 Build the Semantic Layer (Business Logic)

1. Back to DBT >> Studio >> File Explorer
2. Create folder: models/business/
3. Create model file: models/business/fact\_trips.sql
4. Copy the following code into the model.

-- models/business/fact\_trips.sql  
{{ config(schema='mart', materialized='table') }}  
  
SELECT  
 t."lpep\_pickup\_datetime"::DATE AS trip\_date,  
 t."vendor\_id",  
 t."passenger\_count",  
 t."trip\_distance",  
 t."fare\_amount",  
 t."tip\_amount",  
 t."total\_amount",  
 pu."Zone" AS pickup\_zone,  
 dol."Zone" AS dropoff\_zone  
FROM {{ ref('stg\_green\_tripdata') }} t  
LEFT JOIN {{ ref('stg\_zone\_lookup') }} pu   
 ON t."PULocationID" = pu."LocationID"  
LEFT JOIN {{ ref('stg\_zone\_lookup') }} dol   
 ON t."DOLocationID" = dol."LocationID"  
WHERE t.passenger\_count > 0

# 🔨 Build the Model

* In the CLI, at the bottom of the screen, run dbt build:

dbt build

* Navigate to NeonDB and confirm the new schema and table in Neon exist.
* Explore the new table by running query below:

SELECT \* FROM dbt\_smubashir\_mart.fact\_trips LIMIT 10;

# 📈 Analyze the Data

* In Neon DB >> Navigate to SQL Editor
* Run the following query to get a business result set.

-- Top 10 pickup zones by total revenue  
SELECT pickup\_zone, SUM(total\_amount) AS revenue  
FROM dbt\_smubashir\_mart.fact\_trips  
GROUP BY pickup\_zone  
ORDER BY revenue DESC  
LIMIT 10;

1. For bonus insight run the following query:

***SELECT pickup\_zone, dropoff\_zone, AVG(fare\_amount) AS avg\_fare, AVG(tip\_amount) AS avg\_tip, COUNT(\*) AS num\_trips FROM dbt\_smubashir\_mart.fact\_trips GROUP BY pickup\_zone, dropoff\_zone ORDER BY avg\_tip DESC LIMIT 50;***