DBT

Data Transformation Service

JERIN SAM

Table of Contents

[DBT 2](#_Toc184731142)

[What DBT Is 2](#_Toc184731143)

[What DBT Is Not 3](#_Toc184731144)

[How DBT Works 3](#_Toc184731145)

[DBT Helps To 3](#_Toc184731146)

[Setup DBT and Configure DBT Project 4](#_Toc184731147)

[Setup DBT 4](#_Toc184731148)

[Configure and create DBT project 5](#_Toc184731149)

[DBT Components 6](#_Toc184731150)

[Models 6](#_Toc184731151)

[Implementation 6](#_Toc184731152)

[Materialization 7](#_Toc184731153)

[Types of Materialization: 7](#_Toc184731154)

[Implementation 8](#_Toc184731155)

[Seeds 13](#_Toc184731156)

[Implementation 13](#_Toc184731157)

[Sources 14](#_Toc184731158)

[Source Freshness 15](#_Toc184731159)

[Implementation 15](#_Toc184731160)

[Sources vs Seeds 17](#_Toc184731161)

[Snapshots 17](#_Toc184731162)

[Implementation 18](#_Toc184731163)

[Appendix 18](#_Toc184731164)

[Important Links 18](#_Toc184731165)

# DBT

DBT is an open-source data transformation tool designed for analytics engineering. It is widely used for ***transforming***, ***testing***, and ***documenting data*** in data warehouses. It is the “T” in ETL, it is used for transformation using SQL based syntax.

In DBT, Transformation is ***version controlled***.

A screenshot of a computer

Description automatically generated

DBT is not a processing engine. Instead, it is a data transformation tool that ***relies on the processing power of your data warehouse or query engine***.

## What DBT Is

* **SQL-Based Transformation Tool:** DBT uses SQL to define data transformations. You write SQL queries (models), and DBT compiles them into SQL scripts that your data warehouse executes.
* **Orchestration of SQL Jobs:** DBT organizes, manages, and orchestrates SQL transformations, but it does not directly handle the processing of data.
* **Workflow Management:** DBT handles dependencies between data models, allowing you to build complex workflows. It ensures that transformations run in the correct order based on these dependencies.
* **Code Management:** DBT provides a structured framework for managing SQL code, applying data tests, documenting transformations, and version-controlling data pipelines.

## What DBT Is Not

* **DBT is** **not a Database or Processing Engine**: DBT does not store data, nor does it have its own query execution engine. It depends on a data warehouse, such as:
  + BigQuery
  + Snowflake
  + Redshift
  + PostgreSQL
  + Databricks
  + Azure Synapse
* **Not an ETL Tool:** DBT focuses only on the Transformation (T) aspect of the ETL/ELT process. It does not handle Extraction (E) or Loading (L) of data from source systems into a data warehouse.

## How DBT Works

* **SQL Compilation:** DBT takes your SQL files (models) and Jinja templates, compiles them into raw SQL statements, and submits them to your data warehouse for execution.
* **Data Warehouse Execution:** The actual data processing happens in the data warehouse. DBT does not perform the heavy lifting; it delegates the query execution to the warehouse.
* **Results Storage:** Once the data warehouse completes the query, the results are stored in the tables or views specified by DBT.

## DBT Helps To

1. Modelling changes are easy to follow and revert by using version control feature of DBT.
2. View and create dependencies between models.
3. Data quality checks.
4. Error reporting.
5. Incremental load of fact tables.
6. Track history of dimension tables.
7. Easy-to-access documentation.

# Setup DBT and Configure DBT Project

## Setup DBT

***Refer:***

* ***Dockerfile -*** */dbt-learn-hands-on/install\_and\_config/2-dbt-install/Dockerfile*
* ***Python Package Requirement*** - */dbt-learn-hands-on/install\_and\_config/2-dbt-install/* *requirements.txt*
* ***Installation Details*** - */dbt-learn-hands-on/install\_and\_config/2-dbt-install/* *script-dockerfile-execute.md*
* ***SQL Server ODBC Installation -*** */dbt-learn-hands-on/install\_and\_config/3-install-sql-server-odbc-docker-container/install-odbc-driver-container.md*
* ***Host Machine VS Code Connection with Container’s Python Interpreter –***

*/dbt-learn-hands-on/install\_and\_config/4-connect-vscode-docker-container/config-vscode-connect-python.md*

In this tutorial, DBT is installed on docker container. Following are the steps –

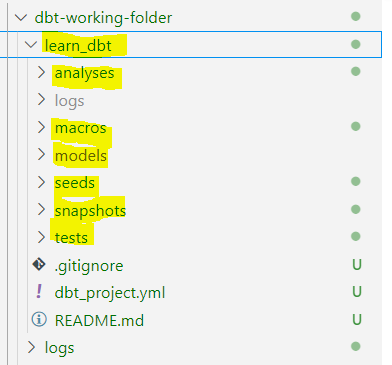
1. Create a *Dockerfile*, which contains Python 3.11.10 slim bookwork image.
   * Types of Docker Images
     + Bookworm - refers to the codename for the latest stable release of Debian.
     + Slim - indicates a minimal version of a Debian distribution with only the essential packages installed
     + Bullseye - is the codename for the previous stable Debian release
     + Alpine - is the “Dockerized” version of Alpine Linux
2. Install *dbt-core* and *dbt-sqlserver* *adapter* via *Dockerfile*.
3. Create working directory in host machine “***/dbt-learn-hands-on/main/dbt-working-folder***” and mount it with Docker container’s working directory using *Dockerfile*, so that any changes done through host machine can be synced in container’s working directory and vice versa.
4. Manually install SQL Server ODBC driver in the container.
5. After the ODBC setup, connect host machine’s VS Code to python interpreter present in Docker container.

## Configure and create DBT project

***Refer: /dbt-learn-hands-on/main/1-configure-dbt-project.md***

Following steps are taken to configure and create DBT folder structure in the container –

1. Create a Python Virtual Environment in the Docker Container and Activate It.
2. Create DBT Config/Profile Folder – Profile folder ***.dbt*** should be created at ***root*** directory. While initializing a DBT project using ***dbt init <project\_name>***a ***profiles.yml*** file will be created at this location to hold connection details of the database to be used by the dbt for the transformations. In this case its SQL Server.
3. Initialize DBT Project and Configure Database using ***dbt init <project\_name>*** - after executing the initialization script, a dbt folder structure will be created.



1. Test DBT Database Connection using ***dbt debug*** –
   * This command will be executed in the dbt project folder to test the connection with database used by DBT for the transformation, in this case its SQL Server.
   * This command uses ***profiles.yml*** file created in the above step.
   * ***Note -*** Generally, error occurs dung the testing, make sure that SQL Server connection details in the ***profiles.yml*** file should be inline with the reference profile file exist at ***“/main/dbt-working-folder/dbt-sql-server-profile.yml”,*** if required copy the same content in ***profiles.yml*** file.
2. Clean DBT Project Folder – This is required to delete examples exist in the dbt project folder.

# DBT Components

## Models

Models are the basic building blocks to implement the business logics. DBT models are SQL files that contain SQL queries used to define transformations i.e. In DBT, Models folder contains SQL files.

Models are materialized as views, tables etc. this means, it corresponds to a view or table in the data warehouse.

***Materialization*** *refers to how a model's query results are stored in the data warehouse. Models can be* ***materialized as views, tables, or other options*** *depending on the configuration and requirements.*

Models can refer other models and can also use templates and macros.

### Implementation

#### Dataset

Airbnb data is used in this tutorial. Refer: ***/dbt-learn-hands-on/main/2-get-data-for-tutorial.md*** for the details on the dataset.

***Dataset Reference:*** */dbt-learn-hands-on/main/2-get-data-for-tutorial.md*

#### Create Models

***DBT Model Reference:*** */dbt-learn-hands-on/main/dbt-working-folder/learn\_dbt/models/src*

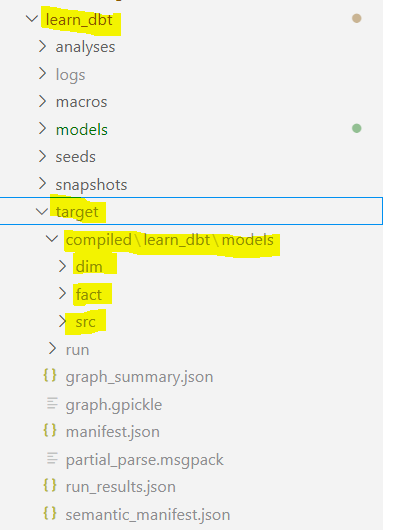
1. Create 3 models in ***/models/src*** folder of the dbt project “***learn\_dbt***”
   1. ***src\_listing.sql*** – This is used to read data from the table ***raw.listings*** and materialized as view ***dbo.listings***
   2. ***src\_reviews.sql*** – This is used to read data from the table ***raw.reviews*** and materialized as view ***dbo.reviews***
   3. ***src\_calendar.sql*** – This is used to read data from the table ***raw.calendar*** and materialized as view ***dbo.calendar***
2. Execute all the created models using ***dbt run.*** This command will create all the necessary DB objects in the database.

#### Debugging

During model execution process, DBT compiles the model sql files and save it at ***/target/compiled*** folder of dbt project folder ***“learn\_dbt”.***

While executing ***dbt run,*** if error occurs then go to ***/target/compiled*** folder in dbt project ***“learn\_dbt”*** and check the compiled sql query of the model, then copy it and execute that query in the database to understand the error.

***dbt compile*** command can also be used to generate the compiled query which can be accessible from ***/target/compiled*** folder



## Materialization

**Materialization** is the process of defining how ***dbt*** should build and store a model (a SQL query) in the database.

Materializations control whether the model's output is stored as a table, a view, or something else.

### Types of Materialization:

#### Table –

* ***What it does:***
  + Executes the query defined in the model and stores the results in a physical table in the database.
* ***When to use:***
  + When the data doesn’t change often.
  + When you need quick access to pre-computed results for performance reasons.
  + Ideal for intermediate or final datasets used in reporting.

1. View –

* ***What it does:***
  + Creates a database view that runs the query in real-time whenever the view is accessed.
* ***When to use:***
  + When the underlying data changes frequently and real-time results are needed.
  + When storage space is a concern (as views don’t store data).
  + **Drawback**: Slower query performance compared to tables since the query is executed every time the view is accessed.

1. Incremental –

* ***What it does:***
  + Adds only new or updated data to an existing table instead of rebuilding the entire dataset.
* ***When to use:***
  + When working with large datasets.
  + When data ingestion or refresh should be efficient and scalable.
* ***Key setup:***
  + An ***is\_incremental()*** macro to define incremental logic.
  + A unique key or timestamp to identify new data.

1. Ephemeral–

* ***What it does:***
  + Executes the query during the build process without materializing it in the database.
* ***When to use:***
  + When the model is only used as an intermediate step in downstream queries.
  + Useful for temporary transformations or intermediate logic.
  + ***Drawback***: Cannot be queried directly in the database.

### Implementation

During the implementation, additional models are created, in these models jinja templates are used to –

* Refer one models in another model.
* Implementation of Materialization.

***Steps:***

1. Create 2 models in ***/models/dim*** folder of the dbt project “***learn\_dbt***”
   1. ***dim\_listing.sql*** – Create dimension table for listings dataset by referencing source model at ***“/models/src***” and use sub-folder level materialization, which will be defined at ***dbt\_project.yml*** file (implementation explanation is given below).
   2. ***dim\_host.sql*** – Create dimension table for hosts dataset by referencing source model at ***“/models/src***” and use sub-folder level materialization, which will be defined at ***dbt\_project.yml*** file (implementation explanation is given below).
   3. ***dim\_listing\_and\_host.sql*** – Create a combined dimension table by referencing dimension models created in above steps and use sub-folder level materialization, which will be defined at ***dbt\_project.yml*** file (implementation explanation is given below).
2. Create 1 models in ***/models/fact*** folder of the dbt project “***learn\_dbt***”
   1. ***fact\_reviews.sql*** – Incremental Load Fact table for reviews, use model level materialization.
3. Update 3 models in ***/models/src*** folder of the dbt project “***learn\_dbt***”
   1. Update all the 3 source models to add model level materialization using jinja templat
4. Use ***dbt run*** to execute the models.
5. Use ***dbt compile*** to compile the models.

#### Model Reference in another Model

During model creation***, jinja templates*** can be used to refer other models as well.

***Syntax: {{ref(‘filename’)}}***

***{{}} this is the jinja template, it is used to tell DBT to dynamically change the content during the compilation.***

A screen shot of a computer

Description automatically generated

#### Materialization at global project level and model sub-folder level–

Materialization implementation can be done at global level, at model sub-folder level (where models are stored) and at individual model sql file level.

*In DBT project,* ***dbt\_project.yml*** *file is used to set the default materialization at global project level and at model’s sub-folder level, this folder exists in the DBT project folder –*

A screenshot of a computer

Description automatically generated

*In* ***dbt\_project.yml****, materialization is added as a config under models -*

A screenshot of a computer screen

Description automatically generated

#### Materialization at model level

In the model sql files, Materialization at global project level and sub-folder level can be overridden by using jinja template in the model file itself.

* + ***Reference- Materialization-Incremental****: Using Jinja template, config method is used to override default materialization -*

A screenshot of a computer

Description automatically generated

* + ***Reference- Materialization-Table****: Using Jinja template, config method is used to override default materialization –*

A screenshot of a computer screen

Description automatically generated

#### Defining Incremental load logic

To implement incremental load, Materialization should be incremental and jinja template should be used along with ***is\_incremental() macro to define the incremental load logic***

A screenshot of a computer program

Description automatically generated

During the execution of ***dbt run***, data will be fully loaded to the database table, and now during the next execution onwards, only incremental data will be loaded to the database table.

If data needs to be fully loaded again then ***dbt run --full-refresh*** can be used.

## 

## Seeds

Seeds are CSV files that are part of your DBT project and can be used to load static or reference data into the database.

**Key Features of Seeds:**

* ***Static Data:*** Seeds are useful for loading small, static datasets, such as lookup tables, mappings, or default configurations.
* ***Version Controlled:*** Since seeds are stored as CSV files in dbt project directory, they can be managed using version control systems like Git.
* ***Materialized as Tables:*** Executing command ***dbt seed***, the CSV files are uploaded to the database and materialized as tables.

**How to Use Seeds:**

Uploading CSV data to ***/seed/*** folder of ***dbt project folder "learn\_dbt"*** and then executing the command ***“dbt seed”*** will push the data to the database.

### Implementation

In this implementation, a onetime load data will be uploaded to the DB.

***Steps:***

1. A csv file will be added to the **/seed/** folder of dbt project folder **"learn\_dbt".**

***Refer*** ***data at location***: /main/dbt-working-folder/learn\_dbt/seeds/seed\_full\_moon\_dates.csv

A screenshot of a computer

Description automatically generated

1. Execute ***dbt seed*** to push the data to the database, which will be materialized as a table in the database.

A screenshot of a computer

Description automatically generated

## Sources

Sources represent tables or views that exist outside of dbt's control, typically raw data in the data warehouse that is loaded by an external process like ETL.

**Key Features of Sources:**

* ***External Data:*** Used to document and query raw or external datasets.
* ***Data Lineage:*** Helps track data lineage from raw sources to transformed outputs.
* ***Tests and Documentation:*** Tests can be applied (e.g., uniqueness, not null) and generate documentation for sources.

**How to Use Sources:**

* Add source definitions in a .yml file (***sources.yml***) inside the ***/models/***directory ***dbt project folder "learn\_dbt"***
* Use the ***source*** function in dbt models to reference sources.

### Source Freshness

***Source Freshness*** refers to the ability to monitor and ensure the timeliness of raw data sources in the database/ data warehouse.

This feature allows to check if your data sources (e.g., external tables or views) are up-to-date based on specified freshness criteria, such as how recently the data was updated.

**How Source Freshness Works:**

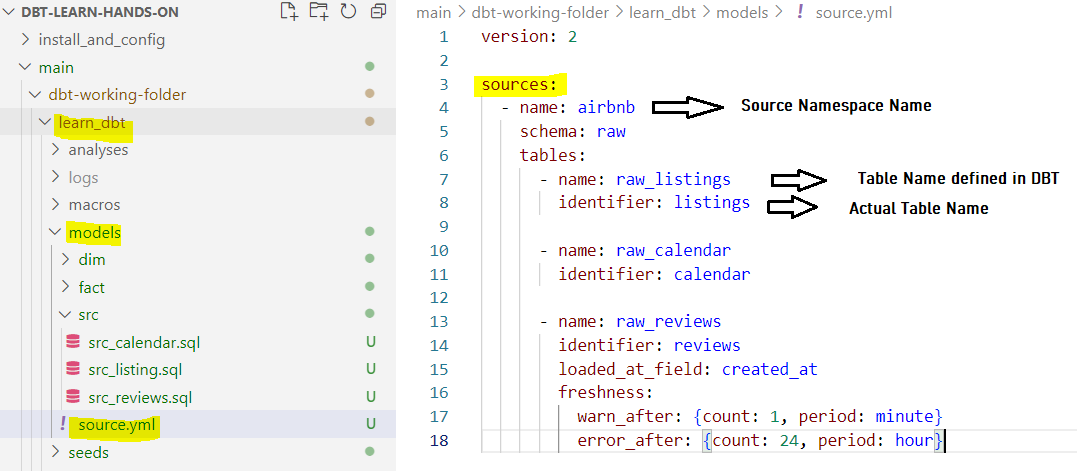
* **Define Freshness Criteria:** Specify freshness rules for your sources in the sources section of a .yml file.
  + The rules include:
    - **loaded\_at\_field:** The column that contains the timestamp of when the data was last updated.
    - **Freshness thresholds:** 
      * **warn\_after:** Time after which a warning will be issued if the data is not fresh.
      * **error\_after:** Time after which an error will be issued if the data is not fresh.
* **Run Source Freshness Checks:** Use ***dbt source freshness*** command to check whether the data meets the defined criteria.
* **View Results:** The results show whether each source is fresh, stale, or failing based on the thresholds you defined.

### Implementation

In this implementation, Sources yaml file will be created in ***/models/*** folder and then the models in /models/src folder will be updated to add reference from Source. yaml file.

***Steps***:

1. Create ***sources.yml*** file in ***/models/*** folder and add all the raw data tables in it.



1. Update models in ***/models/src*** folder to use sources mentioned in ***sources.yml*** file.

A screenshot of a computer

Description automatically generated

1. Add source freshness criteria in reviews dataset.

A screenshot of a computer program

Description automatically generated

1. Test the updated models by executing ***dbt run*** or ***dbt compile***
2. Check Source Freshness by executing ***dbt source freshness***

***Received below Error since data is stale beyond 24 hours as configured in sources.yml file***

A screenshot of a computer program

Description automatically generated

## Sources vs Seeds

|  |  |  |
| --- | --- | --- |
| Feature | Seeds | Sources |
| Definition | Local CSV files in the dbt project. | External tables/views in the database. |
| Purpose | Load static/reference data. | Access raw/external data. |
| Reference Method | {{ ref('seed\_name') }} | {{ source('source\_name', 'table\_name') }} |
| Location | Stored in /seed/ folder in the project. | Already present in the database. |
| Materialization | Always materialized as tables. | Materialized externally. |

## Snapshots

***Snapshot*** is a feature that allows to track changes in data over time by capturing and storing the historical state of a table or query results.

Snapshots are particularly useful for slowly changing dimensions (SCDs) or for maintaining a history of changes to records in your data warehouse.

**How Snapshots Work:**

1. ***Define a Snapshot***:
   1. Create a snapshot configuration in ***/snapshot/*** folder of ***dbt project “learn\_dbt”***, specifying how changes should be detected in data. Jinja template is used to add the configuration.
   2. Snapshots track changes by comparing the current state of a record with its previous state.
2. ***Materialize Changes:***
   1. Executing command ***dbt snapshot***, dbt creates or updates a table in the database that stores the historical records.
3. ***Store Historical Data:***
   1. Snapshots append new rows when changes are detected, allowing to maintain a historical log of data changes.

**Snapshot Strategies:**

1. ***Timestamp Strategy:***
   1. Uses a timestamp and unique identifier column (***e.g., customer\_id,*** ***updated\_at***) to detect changes.
   2. Best for tables that have a clear last-updated timestamp.
2. ***Check Strategy:***
   1. Compares specified columns (***check\_cols***) to detect changes.
   2. Useful for tables without a timestamp column.

### Implementation

In this implementation, SCD Type 2 will be implemented on raw listings dataset.

**Steps:**

1. Create a snapshot sql file for listing dataset (reference from sources.yml file) in ***/snapshot/*** folder of dbt project ***learn\_dbt***

A screenshot of a computer program

Description automatically generated

1. Execute command ***dbt snapshot*** to create a snapshot table, which will create additional columns to support SCD Type 2.
2. Update main raw listings table using below sql command and then re-execute ***dbt snapshot***

select minimum\_nights from [raw].[listings]

WHERE ID = '25840298'

UPDATE [raw].[listings]

SET minimum\_nights = 40, updated\_at = GETDATE()

WHERE ID = '25840298'

1. Check SCD Type 2 in snapshot table created by dbt.

A screenshot of a computer

Description automatically generated

# Appendix

## Important Links

* Dockerfile

<https://docs.docker.com/get-started/docker-concepts/building-images/writing-a-dockerfile/>

* Dockerfile CMD, RUN and ENTRYPOINT

<https://www.docker.com/blog/docker-best-practices-choosing-between-run-cmd-and-entrypoint/>

* Docker mount

<https://docs.docker.com/engine/storage/bind-mounts/>

* DBT SQL Server Installation

<https://docs.getdbt.com/docs/core/connect-data-platform/mssql-setup>