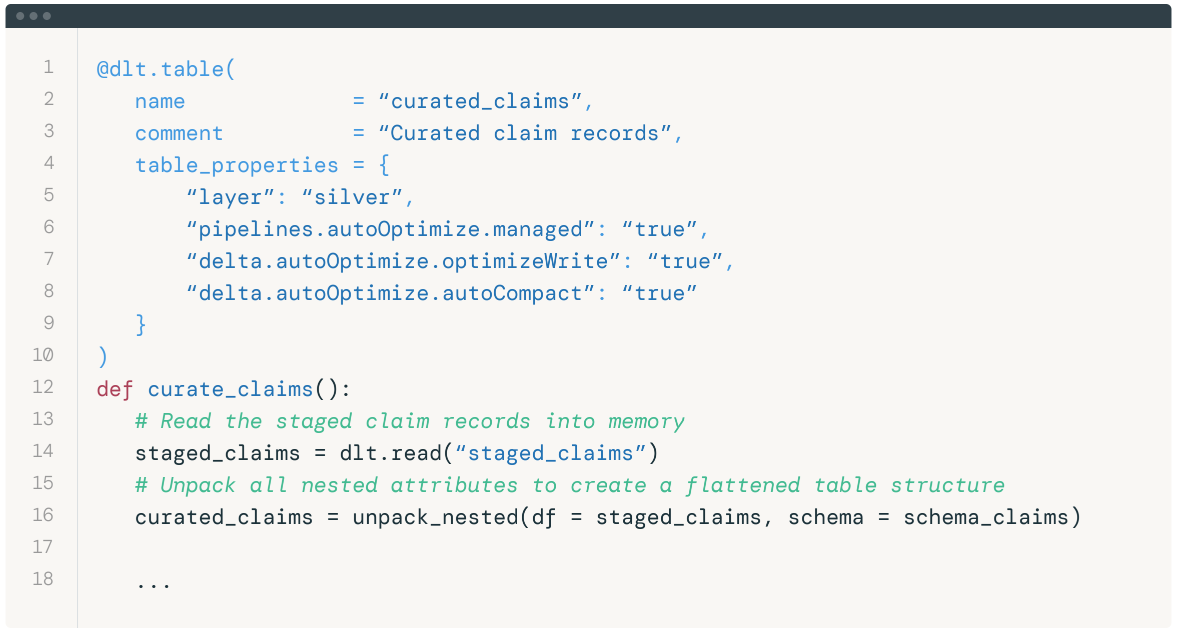
It is important to note that if the source data contains semi-structured or unstructured values, those attributes will be flattened during the conversion process. This means that the results will be stored in grouped text-type columns, and these entities will have to be dissected and unpacked with DLT in the curation process to create separate attributes.

Step 2: Automating the workflow

With the data in the lakehouse, we can use Delta Live Tables (DLT) to build a simple, automated data engineering workflow. DLT provides a declarative framework for specifying detailed feature engineering steps. Currently, DLT supports APIs for both Python and SQL. In this example, we will use Python APIs to build our workflow.

The most fundamental construct in DLT is the definition of a table. DLT interrogates all table definitions to create a comprehensive workflow for how data should be processed. For instance, in Python, tables are created using function definitions and the `dlt.table` decorator (see example of Python code below). The decorator is used to specify the name of the resulting table, a descriptive comment explaining the purpose of the table, and a collection of table properties.



```
@dlt.table(
   name             = “curated_claims”,
   comment          = “Curated claim records”,
   table_properties = {
       “layer”: “silver”,
       “pipelines.autoOptimize.managed”: “true”,
       “delta.autoOptimize.optimizeWrite”: “true”,
       “delta.autoOptimize.autoCompact”: “true”
   }
)
def curate_claims():
   # Read the staged claim records into memory
   staged_claims = dlt.read(“staged_claims”)
   # Unpack all nested attributes to create a flattened table structure
   curated_claims = unpack_nested(df = staged_claims, schema = schema_claims)

   ...
```

Instructions for feature engineering are defined inside the function body using standard PySpark APIs and native Python commands. The following example shows how PySpark joins claims records with data from the policies table to create a single, curated view of claims.