

Short Summary

Last week, we integrated and explored multiple datasets using Python's pandas library, identifying key entities, attributes, and relationships. Data inconsistencies such as mismatched types and null values were noted.

Initially, we defined the business process as "**every time a user places an order,**" but this week we corrected it to "**every product for every order a user makes.**" The focus shifted from order_id to a newly defined **line_id**. The conceptual model was revised accordingly, and Kimball's dimensional modeling was still applied to implement the logical and physical models in SQL.

- We designed the ETL/ELT pipeline in Kestra, creating a Git repository with seeds/ and models/ directories similar to a **tutorial we found of Kestra UI**.
- Source files are converted to CSV with python
- Using *io.kestra.plugin.dbt.cli.DbtCLI* Kestra can find and read the files on our repository.
- CSV files are staged using queries on file names **stg_*.sql** where we rename fields, cast data types, and standardize format.
- Results are transformed according to the schema using queries on file names ***_dim.sql** and ***_fact.sql**.
- Finally, we export the dim and fact tables as *outputFiles* on *Kestra UI* using python.

Before we move on, we need consultation to finalize project decisions, including confirming assumptions, where to store output, task modularization, removal of hardcoding, and whether datasets will be updated.

Evidence of progress

```

4   tasks:
5     - id: dbt
6     tasks:
7       - id: dbt.build
8         profiles: |
9
10
11       - id: export
12         type: io.kestra.plugin.scripts.python.Script
13         outputFiles:
14           - ".csv"
15         taskRunner:
16           type: io.kestra.plugin.scripts.runner.docker.Docker
17           containerImage: ghcr.io/kestra-io/duckdb:latest
18           script: |
19             import duckdb
20             import pandas as pd
21
22             conn = duckdb.connect(database='dbt.duckdb', read_only=False)
23
24             tables_query = "SELECT table_name FROM information_schema.
25             tables WHERE table_schema = 'main';"
26
27             tables = conn.execute(tables_query).fetchall()
28
29             # Export each table to CSV, only if the table name ends with
30             # 'dim' or 'fact'
31             for table_name in tables:
32               table_name = table_name[0]
33               if table_name.endswith('dim') or table_name.endswith
34                 ('fact'):
35                 query = f"SELECT * FROM {table_name}"
36                 df = conn.execute(query).fetchdf()
37                 df.to_csv(f"{table_name}.csv", index=False)
38
39             conn.close()

```

File	Commit Message	Date
models	Create staff_dim.sql	yesterday
seeds	Delete seeds/schema.yml	yesterday
dbt_project.yml	Create dbt_project.yml	3 days ago
profiles.yml	Create profiles.yml	3 days ago

staff_id	staff_name	staff_job_level	staff_street	staff_state	staff_city	staff_country	staff_c
STAF009956	Randall Bergstrom	Intermediate	376 Land chester	Texas	Omaha	Cod	NC
STAF009956	Christian Hessel	Intermediate	945 West Camp shire	New Mexico	Albuquerque	City	1488
STAF001519	Jordi Gleicher	entry	720 Centers burgh	Virginia	Bakersfield	Timon-Leste	6492588
STAF003616	Price Hintz	Intermediate	1720 North Skyway	Alabama	Scottsdale	Palestine	3281
STAF002683	Garfield Legros	Intermediate	9 Way town	Virginia	San Bernardino	Antarctica	349
STAF0062193	Shanna Keanan	entry	2134 North Pines	Massachusetts	Hampshire	Thailand	36136
STAF0033256	Curt Thiel	entry	8111 Isle Furt	New York	Anahiem	Gambia	2887566959
STAF0045925	Kiley Monahan	Intermediate	2222 Land bury	South Dakota	Indianapolis	Mongolia	115
STAF0029759	Coleman Haag	Intermediate	2669 Row side	North Carolina	San Francisco	North Mace	115
STAF0086683	Jacques Metz	entry	38253 South Ridge mouth	Minnesota	Kansas	South Georgia	the
STAF0082157	Destini Connell	Intermediate	9457 East Ridge	South Carolina	Atlanta	Barbados	5988
STAF0081726	Tevin Weimann	entry	139 North Coves mouth	West Virginia	Scottsdale	Slovakia	7110
STAF0028899	Grady Howell	entry	743 Turnpike side	Maine	Arlington	Eritrea	6858229888
STAF00841366	Zsckery Kris	Intermediate	41981 Port Mill	Nebraska	Irvine	Moldova	2821-61
STAF0083955	Rick Marvin	entry	2899 Branch ing	California	Kansas	Viet Nam	3550941247
STAF0083256	Shawn Conner	Intermediate	1121 Lake	Michigan	Atlanta	Venezuela	203-05
STAF0085683	Sedrick Walter	senior	8293 Wall	Nevada	Toledo	American Samoa	1853723
STAF00948489	Willia Dicki	entry	99372 East Dam chester	Minnesota	Scottsdale	Djibouti	12720156
STAF0062223	Citlalli Runolfsson	Intermediate	97572 Port Creek	Town	Wisconsin	Norfolk	Cabo Ve
STAF0058581	Vernon Mann	senior	36123 New Court	berg	Illinois	Denver	968492646
STAF0046993	Mozelle Weimann	entry	899 North Passage	port	Indiana	Aurora	Macao
STAF0014985	Jennings Johnston	senior	9319 North Lake	bury	Missouri	Toledo	Nigeria