Creating a ETL pipeline

Date: November 15, 2024

Executive Summary

In this assignment, we focused on populating data warehouse tables using the Apache Hop ETL platform. The task involved working with three dimension tables—Product Dimension, Customer Dimension, and Date Dimension and a fact table called Sales. Initially, we created the dimension tables using DDL files provided by the professor. While setting up these tables, we defined sequence indices. The Date Dimension, originally spanning from 2018 to 2026, required modification to meet the client's need for a date range starting in January 2016 and extending to December 2027. To accommodate this, we adjusted the DDL to include dates over 4,018 days, covering the extended range. After this adjustment, we populated the Customer and Product Dimension tables using insert statements.

Once the tables were set up, we prepared the environment for Apache Hop by downloading and unzipping the platform, along with Zulu Oracle JDK 11, to enable the required Java environment. To connect Apache Hop with the Oracle database, we downloaded the Oracle JDBC driver and configured the connection using the wallet credentials from the database. Initially, we attempted to use a "High" performance connection string. However, when creating a pipeline for the Product Dimension, we encountered an error saying, "cannot modify object in parallel after modifying it". Attempts to resolve this through SQL commands to disable parallel processing were unsuccessful. To address this issue, we created a new connection using the "Low" performance string, where parallel processing is disabled by default. This change resolved the error, and we successfully ran the pipeline.

Following this, we proceeded to load data into the Customer and Product Dimensions by carefully defining appropriate Slowly Changing Dimension (SCD) types for the relevant fields. For the Sales Fact Table, we constructed a pipeline with lookups for the CustomerKey and ProductKey, using CustomerID and ProductID as reference points. This pipeline allowed us to successfully load 1,000 rows into the Sales Fact Table.

To ensure the integrity of the data and the success of our pipelines, we conducted quality assurance checks by running SELECT queries on the dimension and fact tables, validating the

results, and capturing screenshots for documentation. Additionally, screenshots of the completed pipelines were included to showcase the workflow.

In conclusion, this assignment provided valuable hands-on experience in populating a data warehouse and resolving real-world challenges in ETL processes. By adapting to errors and implementing solutions, we successfully met client requirements and ensured the accuracy of the data. This work underscored the importance of understanding database connections, pipeline configuration, and data quality assurance in data engineering.

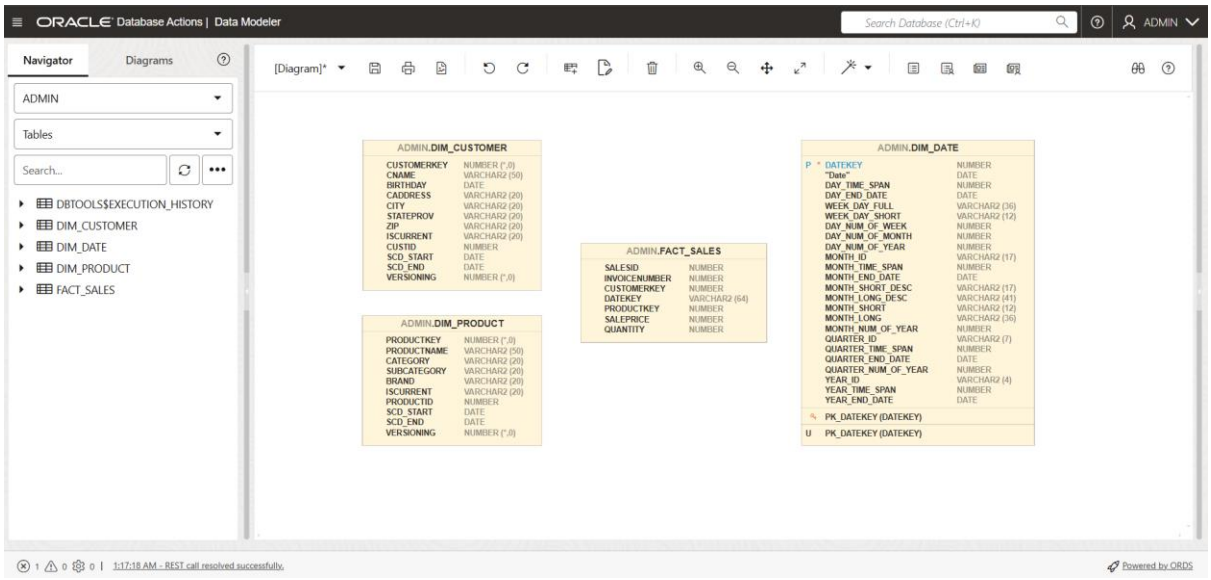
Question 1: Based on the diagram generated, what is this database missing that you'd expect to see? Why might it be missing this component?

The database diagram does not show the connections between the dimension tables and the fact table because primary keys and foreign key relationships were not defined between them. This omission violates the principle of referential integrity, which ensures consistency and valid relationships between the tables.

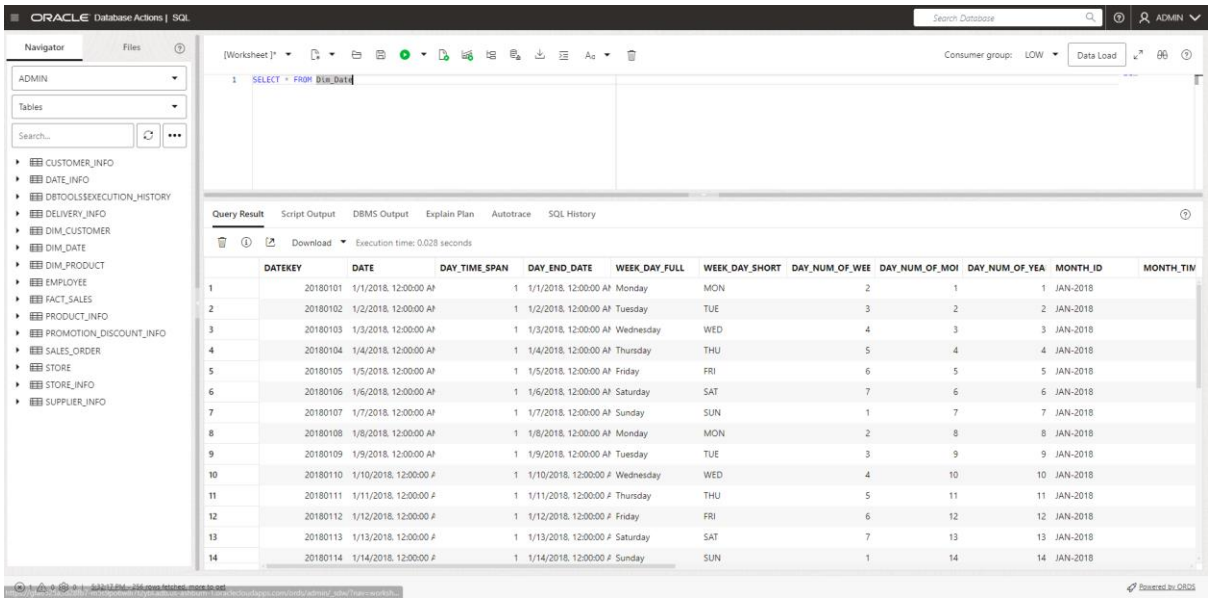
Question 2: You might have noticed we're not doing a lookup for the date dimension. Write 1-2 sentences detailing why we don't need to. You'll be able to figure this out likely by looking at the data in the fact CSV and the date dimension.

Lookups are essential when working with slowly changing dimensions in a database to ensure the correct version or the active version of a record is used during mapping. In the date dimension, there will only be a single version of each row. Therefore, we can directly map the Date Dimension to the fact tables using the `DateKey` field, which is present in both the fact table and the Date Dimension.

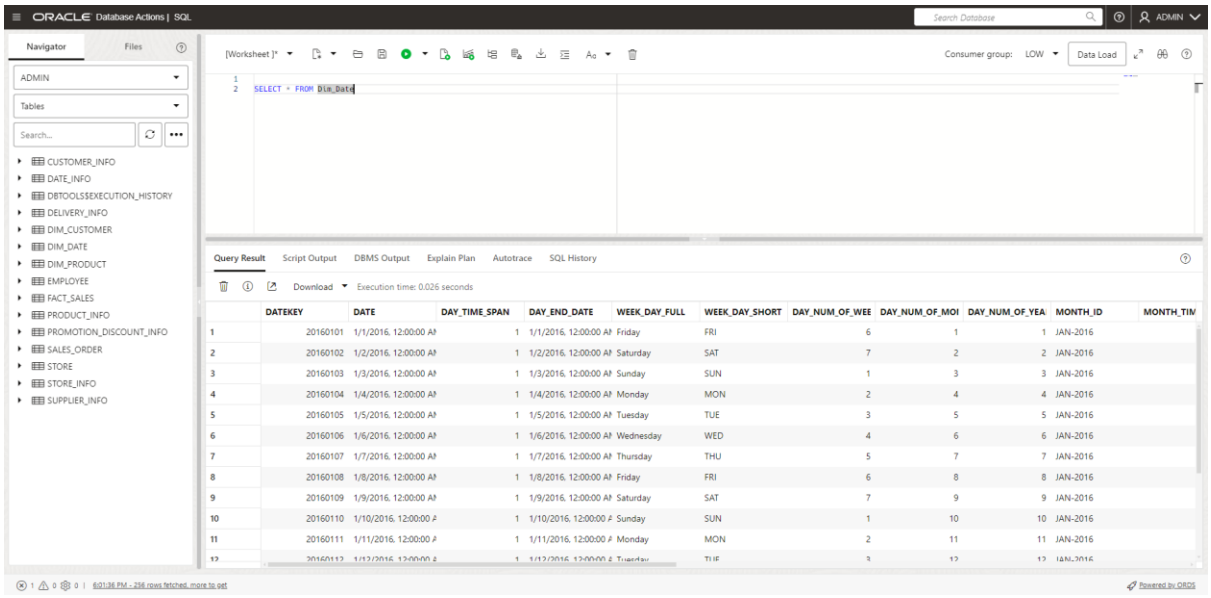
Appendix A: Diagram of database imported.



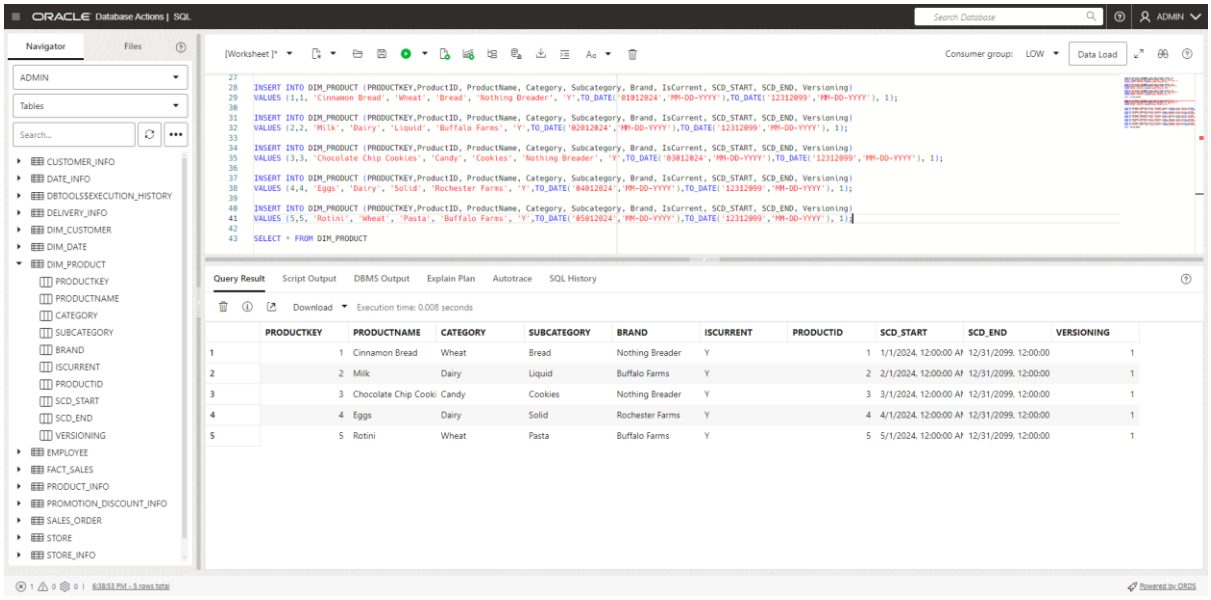
Appendix B: Screenshot of “SELECT * FROM Dim_Date”.



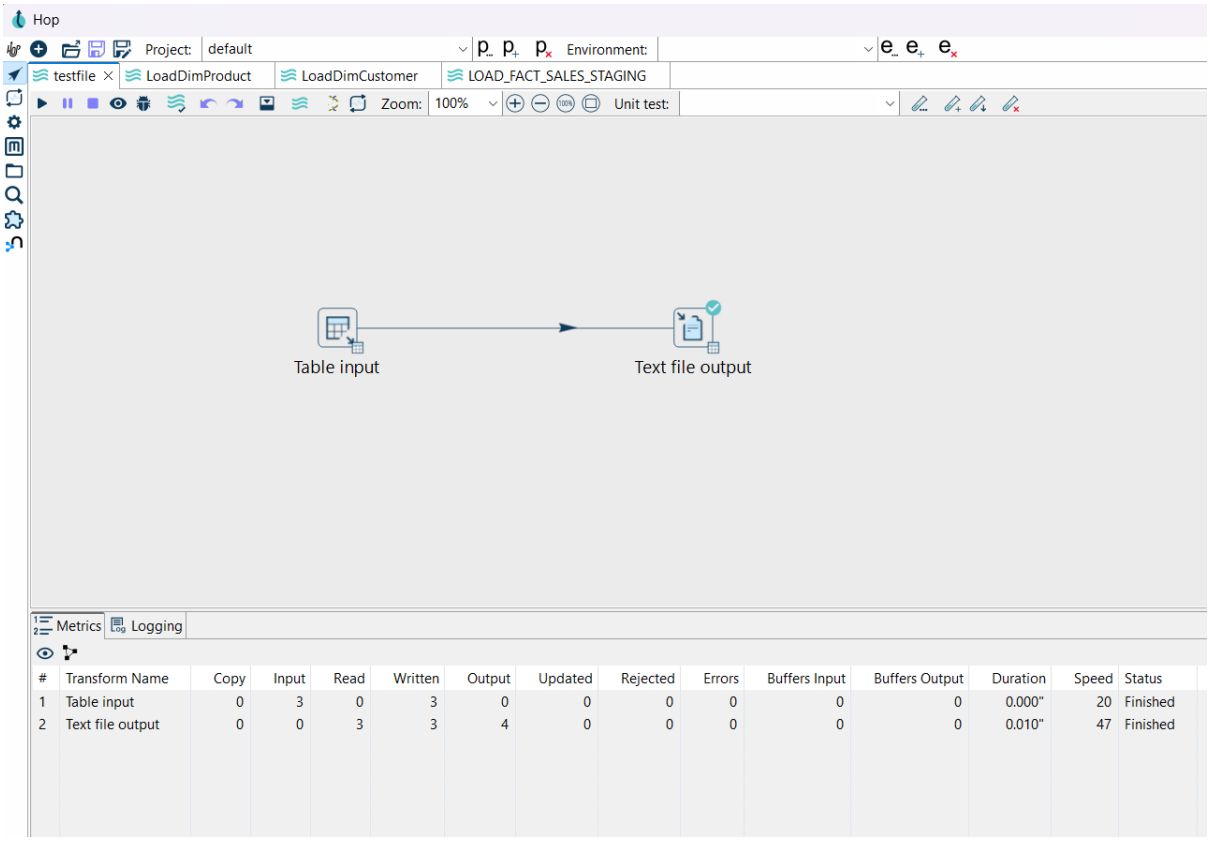
Appendix C: Screenshot of “SELECT * FROM Dim_Date” after updating Dim_Date query.



Appendix D: Screenshot of “Select * from Dim_Product”

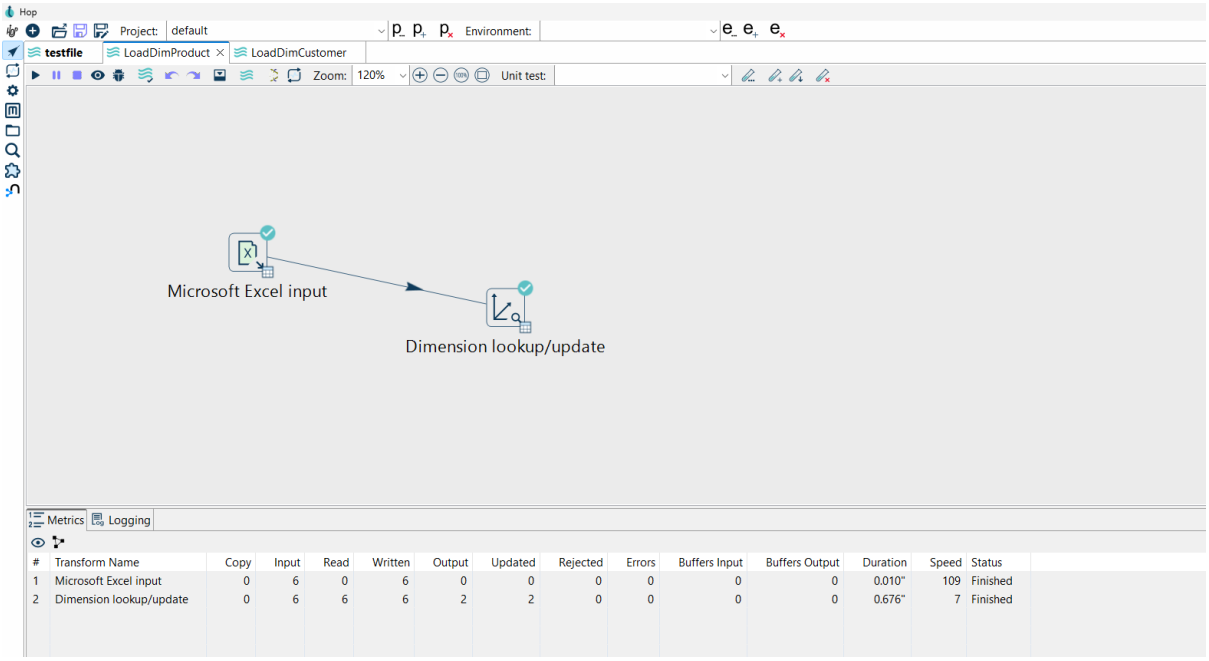


Appendix E: Screenshot of test pipeline.



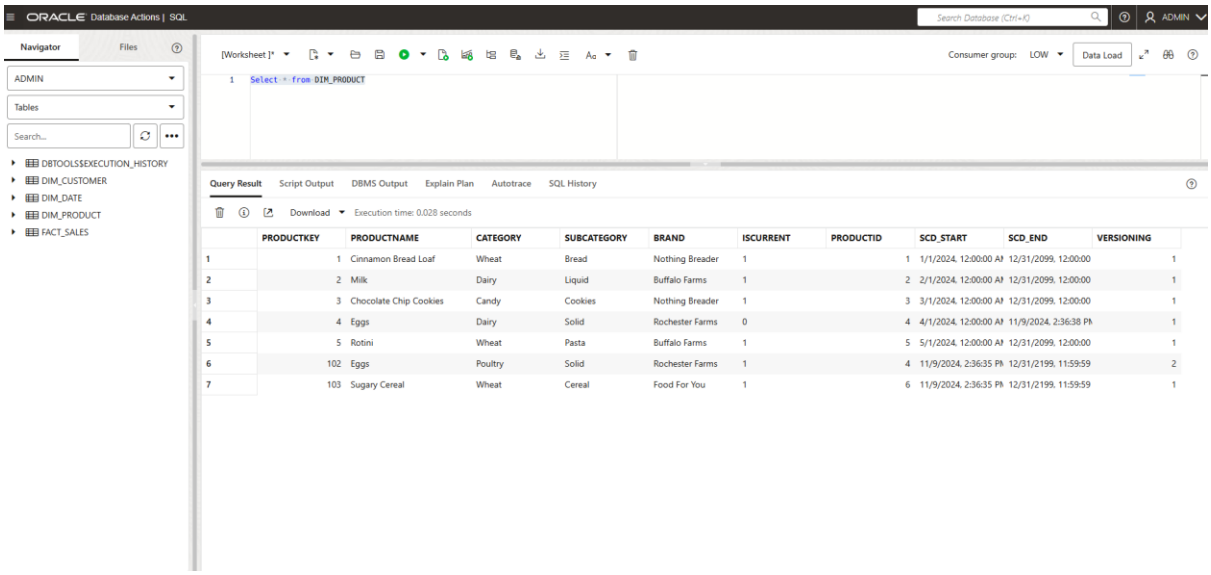
#	Transform Name	Copy	Input	Read	Written	Output	Updated	Rejected	Errors	Buffers Input	Buffers Output	Duration	Speed	Status
1	Table input	0	3	0	3	0	0	0	0	0	0	0.000"	20	Finished
2	Text file output	0	0	3	3	4	0	0	0	0	0	0.010"	47	Finished

Appendix F: Screenshot of pipeline of Dimension lookup for Product dimension



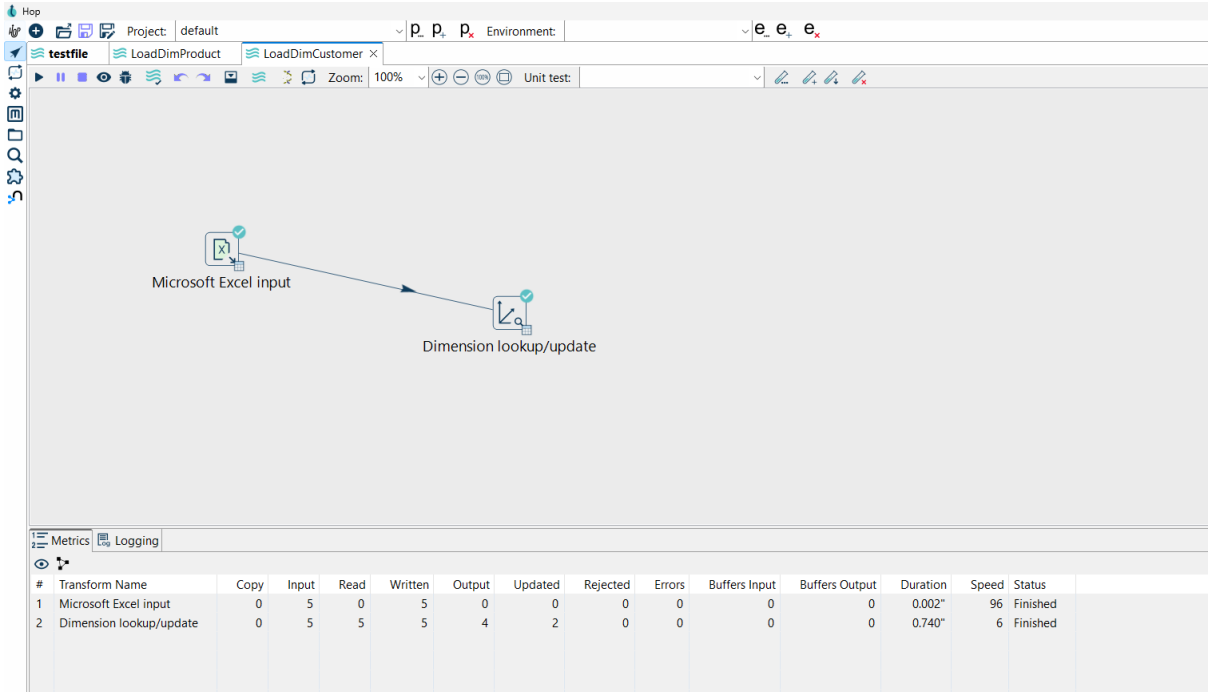
#	Transform Name	Copy	Input	Read	Written	Output	Updated	Rejected	Errors	Buffers Input	Buffers Output	Duration	Speed	Status
1	Microsoft Excel input	0	6	0	6	0	0	0	0	0	0	0.010"	109	Finished
2	Dimension lookup/update	0	6	6	6	2	2	0	0	0	0	0.676"	7	Finished

Appendix G: Screenshot of “Select * from Dim_Product” after running pipeline.



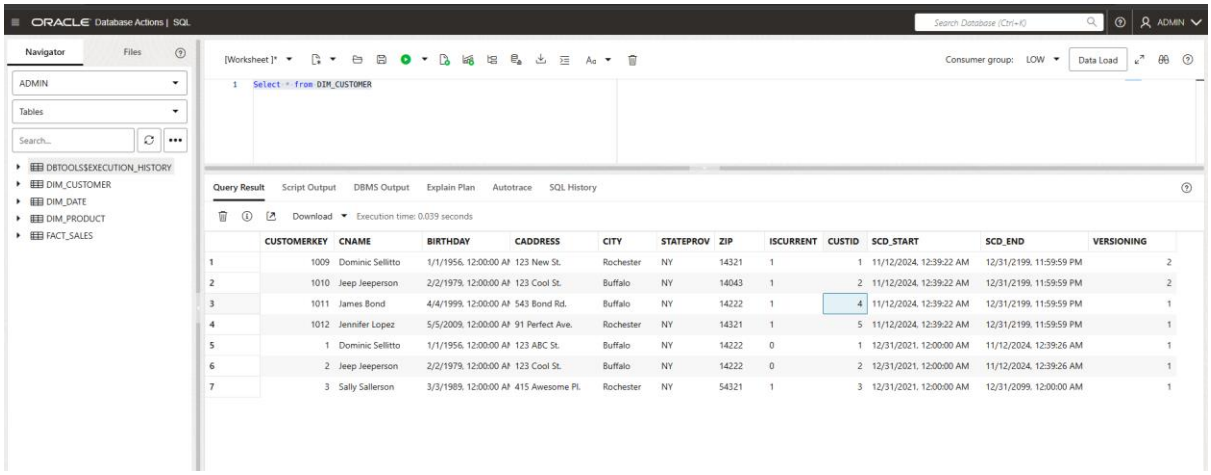
	PRODUCTKEY	PRODUCTNAME	CATEGORY	SUBCATEGORY	BRAND	ISCURRENT	PRODUCTID	SCD_START	SCD_END	VERSIONING
1	1	Cinnamon Bread Loaf	Wheat	Bread	Nothing Breader	1	1	1/1/2024, 12:00:00 AM	12/31/2099, 12:00:00	1
2	2	Milk	Dairy	Liquid	Buffalo Farms	1	2	2/1/2024, 12:00:00 AM	12/31/2099, 12:00:00	1
3	3	Chocolate Chip Cookies	Candy	Cookies	Nothing Breader	1	3	3/1/2024, 12:00:00 AM	12/31/2099, 12:00:00	1
4	4	Eggs	Dairy	Solid	Rochester Farms	0	4	4/1/2024, 12:00:00 AM	11/9/2024, 2:36:38 PM	1
5	5	Rotini	Wheat	Pasta	Buffalo Farms	1	5	5/1/2024, 12:00:00 AM	12/31/2099, 12:00:00	1
6	102	Eggs	Poultry	Solid	Rochester Farms	1	4	11/9/2024, 2:36:35 PM	12/31/2199, 11:59:59	2
7	103	Sugary Cereal	Wheat	Cereal	Food For You	1	6	11/9/2024, 2:36:35 PM	12/31/2199, 11:59:59	1

Appendix H: Screenshot of pipeline of Dimension lookup for Customer dimension



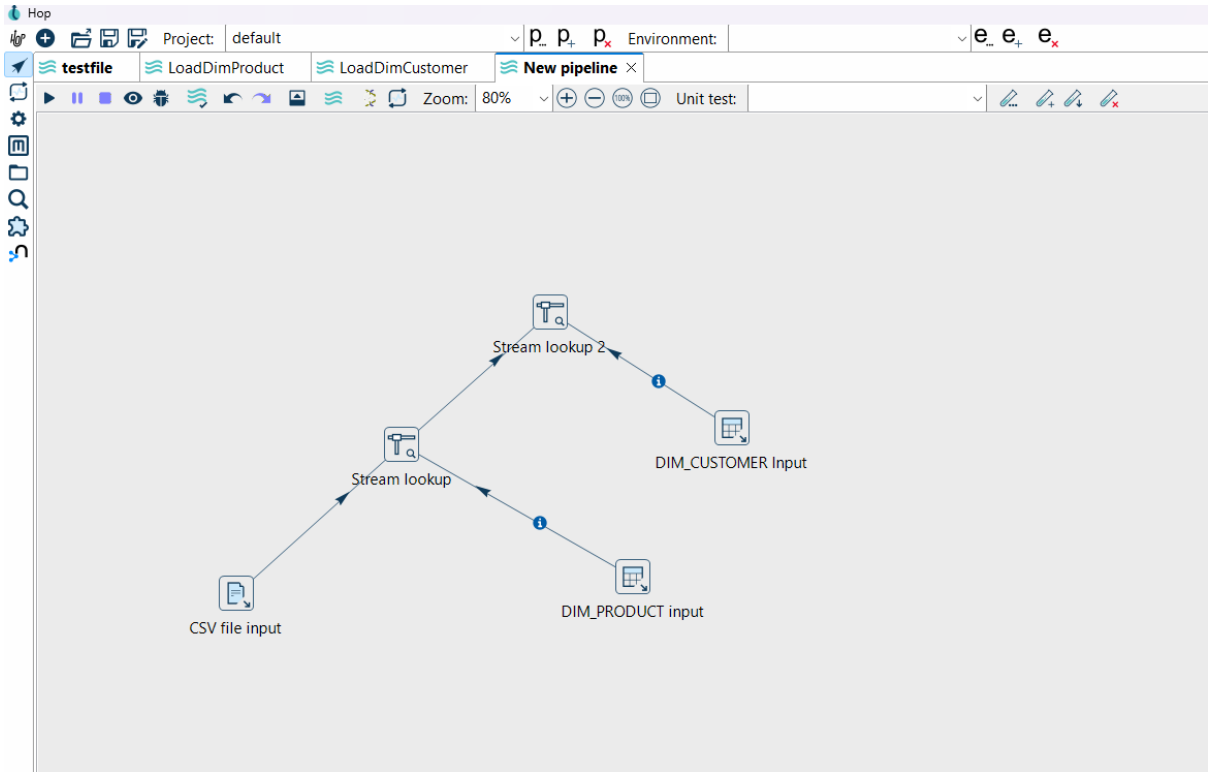
#	Transform Name	Copy	Input	Read	Written	Output	Updated	Rejected	Errors	Buffers Input	Buffers Output	Duration	Speed	Status
1	Microsoft Excel input	0	5	0	5	0	0	0	0	0	0	0.002"	96	Finished
2	Dimension lookup/update	0	5	5	5	4	2	0	0	0	0	0.740"	6	Finished

Appendix I: Screenshot of “Select * from Dim_Customer” after running pipeline.

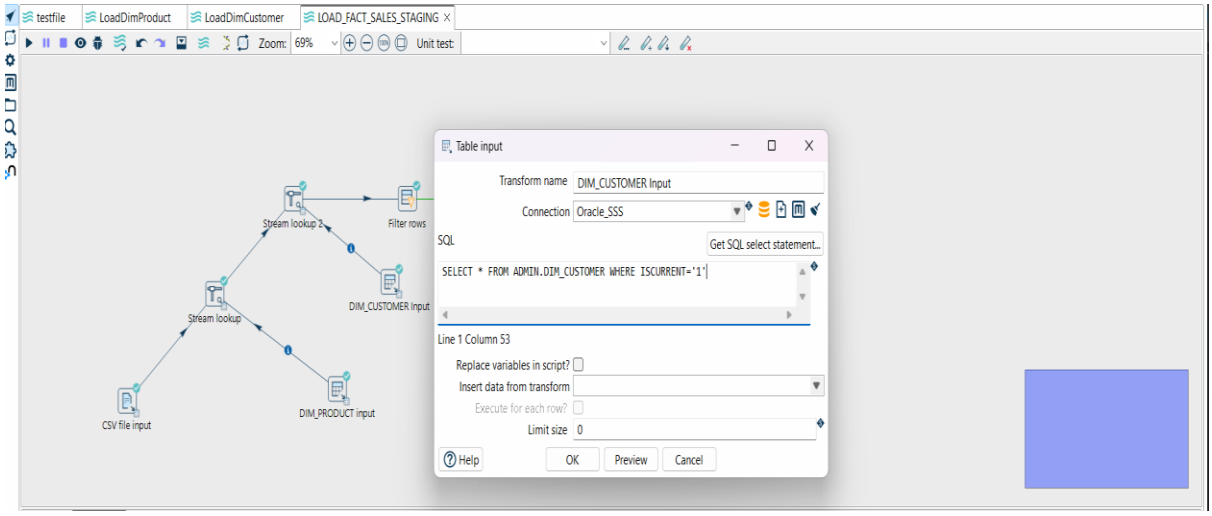


	CUSTOMERKEY	CNAME	BIRTHDAY	CADDRESS	CITY	STATEPROV	ZIP	ISCURRENT	CUSTID	SCD_START	SCD_END	VERSIONING
1	1009	Dominic Sellitto	1/1/1956, 12:00:00 AM	123 New St.	Rochester	NY	14321	1	1	11/12/2024, 12:39:22 AM	12/31/2199, 11:59:59 PM	2
2	1010	Jeep Jeeperson	2/2/1979, 12:00:00 AM	123 Cool St.	Buffalo	NY	14043	1	2	11/12/2024, 12:39:22 AM	12/31/2199, 11:59:59 PM	2
3	1011	James Bond	4/4/1999, 12:00:00 AM	543 Bond Rd.	Buffalo	NY	14222	1	4	11/12/2024, 12:39:22 AM	12/31/2199, 11:59:59 PM	1
4	1012	Jennifer Lopez	5/5/2009, 12:00:00 AM	91 Perfect Ave.	Rochester	NY	14321	1	5	11/12/2024, 12:39:22 AM	12/31/2199, 11:59:59 PM	1
5	1	Dominic Sellitto	1/1/1956, 12:00:00 AM	123 ABC St.	Buffalo	NY	14222	0	1	12/31/2021, 12:00:00 AM	11/12/2024, 12:39:26 AM	1
6	2	Jeep Jeeperson	2/2/1979, 12:00:00 AM	123 Cool St.	Buffalo	NY	14222	0	2	12/31/2021, 12:00:00 AM	11/12/2024, 12:39:26 AM	1
7	3	Sally Sallerson	3/3/1989, 12:00:00 AM	415 Awesome Pl.	Rochester	NY	54321	1	3	12/31/2021, 12:00:00 AM	12/31/2099, 12:00:00 AM	1

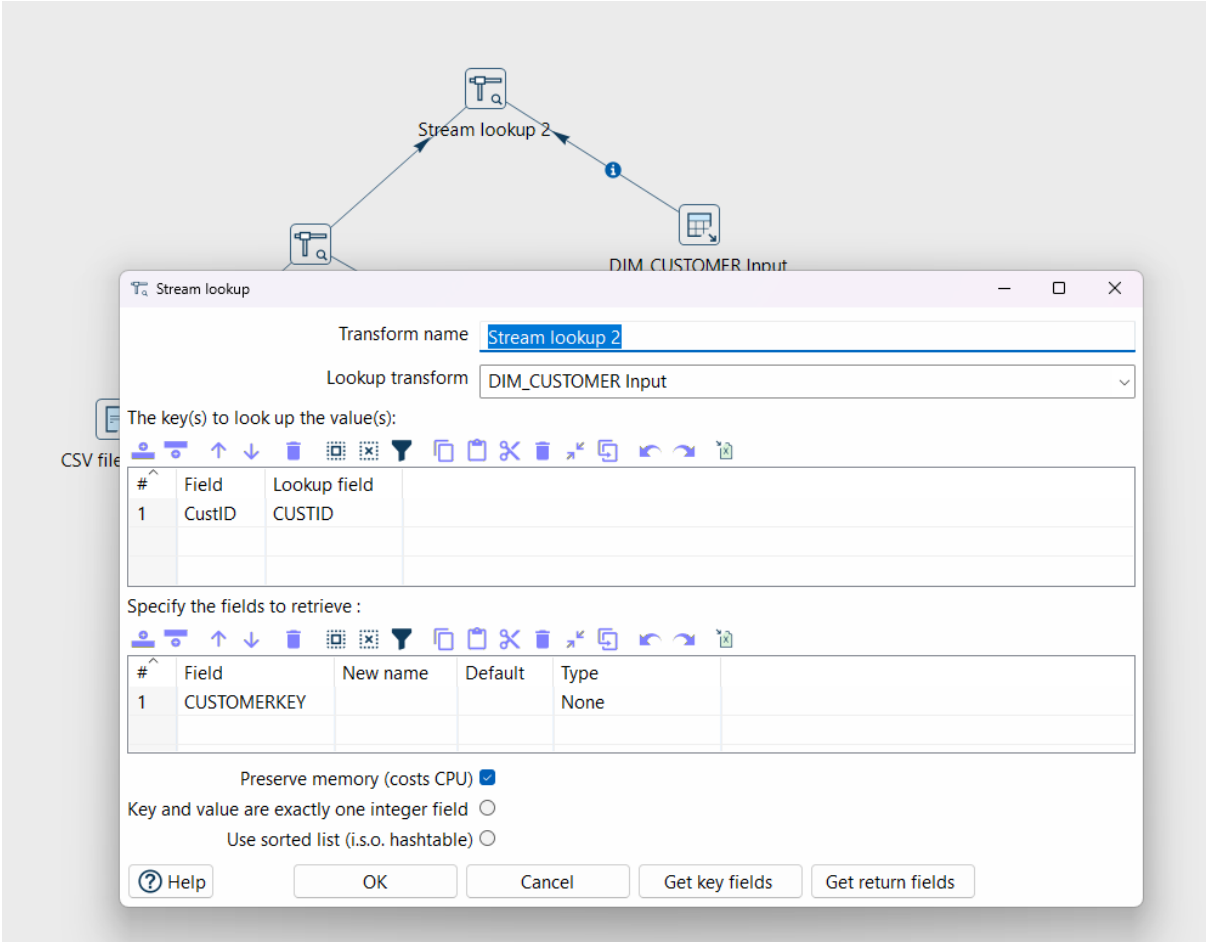
Appendix J: Screenshot of flow of pipeline



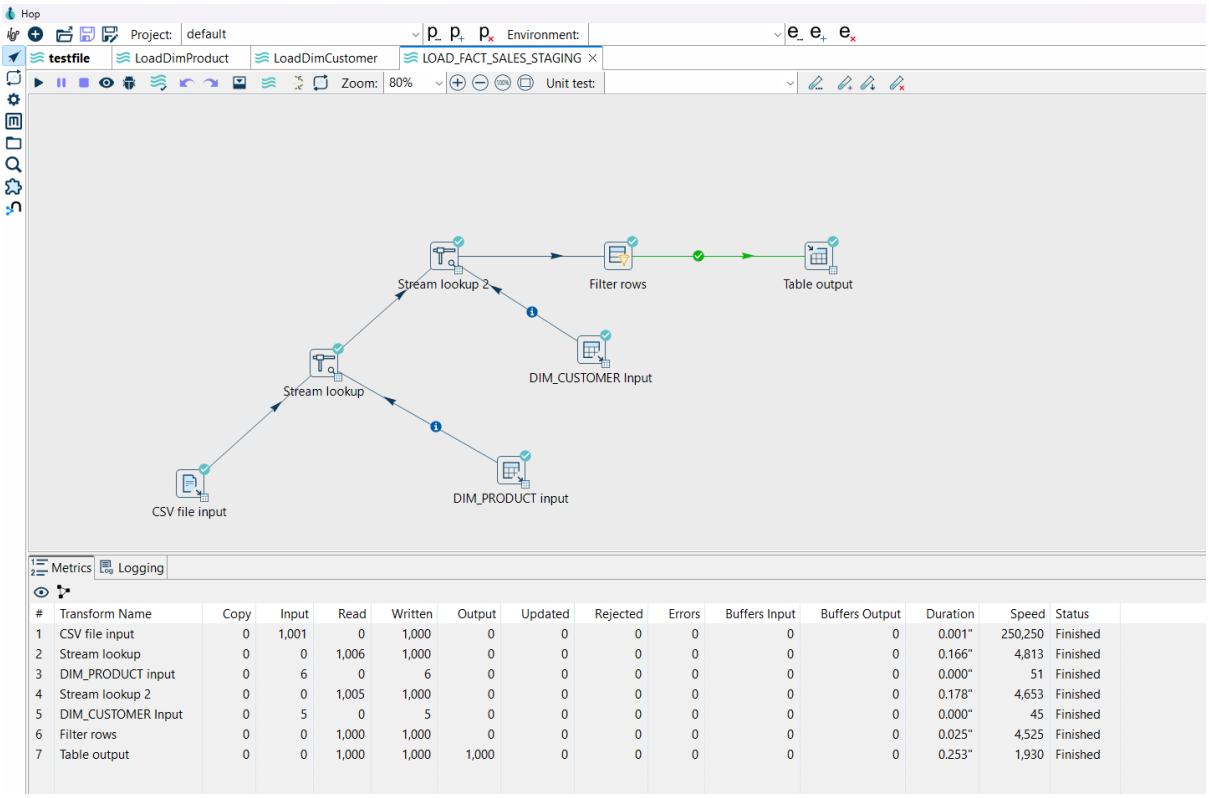
Appendix K: Screenshot of DIM_Customer_input



Appendix L: Screenshot of Stream lookup 2

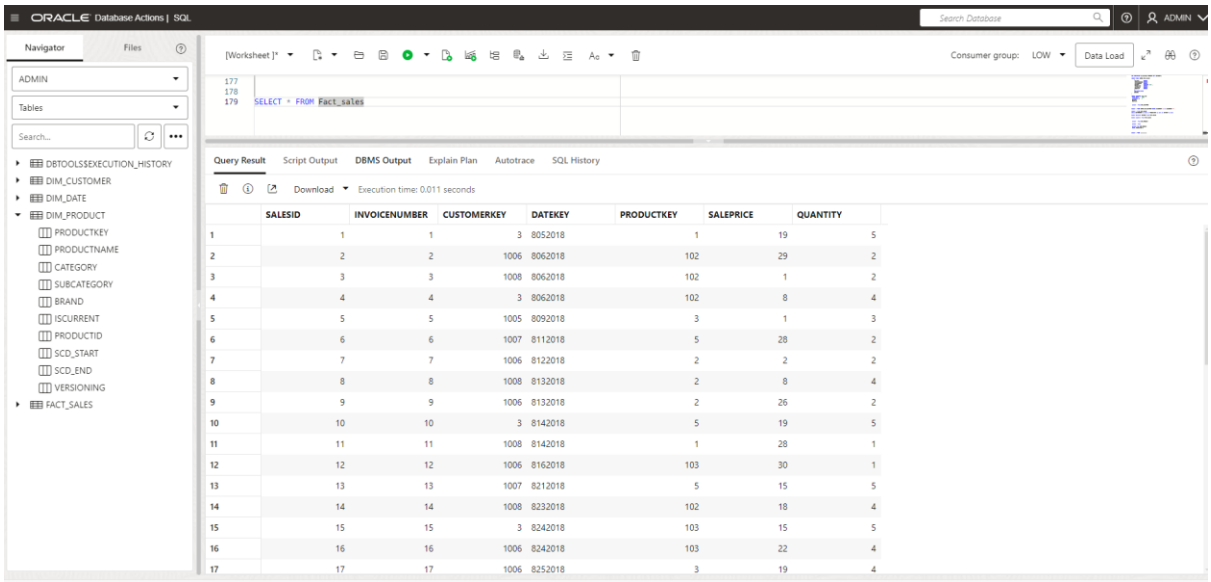


Appendix M: Screenshot of complete pipeline.



#	Transform Name	Copy	Input	Read	Written	Output	Updated	Rejected	Errors	Buffers Input	Buffers Output	Duration	Speed	Status
1	CSV file input	0	1,001	0	1,000	0	0	0	0	0	0	0.001"	250,250	Finished
2	Stream lookup	0	0	1,006	1,000	0	0	0	0	0	0	0.166"	4,813	Finished
3	DIM_PRODUCT input	0	6	0	6	0	0	0	0	0	0	0.000"	51	Finished
4	Stream lookup 2	0	0	1,005	1,000	0	0	0	0	0	0	0.178"	4,653	Finished
5	DIM_CUSTOMER Input	0	5	0	5	0	0	0	0	0	0	0.000"	45	Finished
6	Filter rows	0	0	1,000	1,000	0	0	0	0	0	0	0.025"	4,525	Finished
7	Table output	0	0	1,000	1,000	1,000	0	0	0	0	0	0.253"	1,930	Finished

Appendix N: Screenshot of “Select * from FACT_SALES” after successfully running pipeline.



	SALESID	INVOICENUMBER	CUSTOMERKEY	DATEKEY	PRODUCTKEY	SALEPRICE	QUANTITY
1	1	1	3	8052018	1	19	5
2	2	2	1006	8062018	102	29	2
3	3	3	1008	8062018	102	1	2
4	4	4	3	8062018	102	8	4
5	5	5	1005	8092018	3	1	3
6	6	6	1007	8112018	5	28	2
7	7	7	1006	8122018	2	2	2
8	8	8	1008	8132018	2	8	4
9	9	9	1006	8132018	2	26	2
10	10	10	3	8142018	5	19	5
11	11	11	1008	8142018	1	28	1
12	12	12	1006	8162018	103	30	1
13	13	13	1007	8212018	5	15	5
14	14	14	1008	8232018	102	18	4
15	15	15	3	8242018	103	15	5
16	16	16	1006	8242018	103	22	4
17	17	17	1006	8252018	3	19	4