

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.

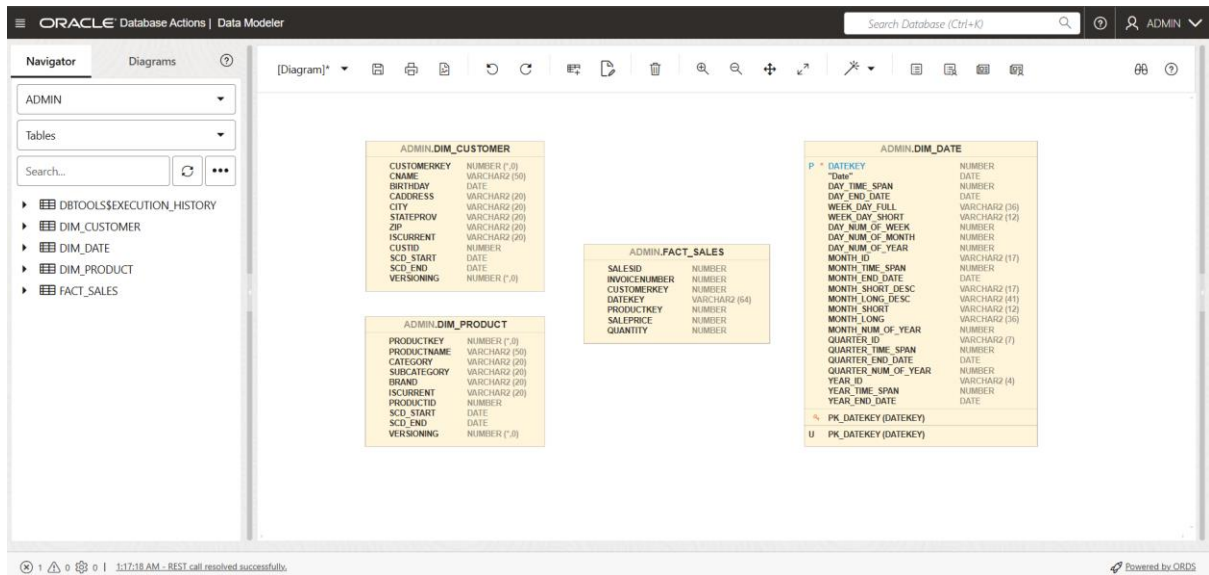
Question 1: Based on the diagram generated, what is this database missing that you'd expect to see? Why might it be missing this component?

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”.

The screenshot shows the Oracle SQL Developer interface. The left sidebar contains a Navigator pane with a tree view showing the database structure: ADMIN, Tables, and a search bar. The main workspace displays a query result for the query "SELECT * FROM Dim_Date". The query result is shown in a table with 14 rows and 11 columns. The columns are: DATEKEY, DATE, DAY_TIME_SPAN, DAY_END_DATE, WEEK_DAY_FULL, WEEK_DAY_SHORT, DAY_NUM_OF_WEEK, DAY_NUM_OF_MONTH, DAY_NUM_OF_YEAR, MONTH_ID, and MONTH_TIME_SPAN. The data is as follows:

	DATEKEY	DATE	DAY_TIME_SPAN	DAY_END_DATE	WEEK_DAY_FULL	WEEK_DAY_SHORT	DAY_NUM_OF_WEEK	DAY_NUM_OF_MONTH	DAY_NUM_OF_YEAR	MONTH_ID	MONTH_TIME_SPAN
1	20180101	1/1/2018, 12:00:00 AM	1	1/1/2018, 12:00:00 AM	Monday	MON	2	1	1	JAN-2018	
2	20180102	1/2/2018, 12:00:00 AM	1	1/2/2018, 12:00:00 AM	Tuesday	TUE	3	2	2	JAN-2018	
3	20180103	1/3/2018, 12:00:00 AM	1	1/3/2018, 12:00:00 AM	Wednesday	WED	4	3	3	JAN-2018	
4	20180104	1/4/2018, 12:00:00 AM	1	1/4/2018, 12:00:00 AM	Thursday	THU	5	4	4	JAN-2018	
5	20180105	1/5/2018, 12:00:00 AM	1	1/5/2018, 12:00:00 AM	Friday	FRI	6	5	5	JAN-2018	
6	20180106	1/6/2018, 12:00:00 AM	1	1/6/2018, 12:00:00 AM	Saturday	SAT	7	6	6	JAN-2018	
7	20180107	1/7/2018, 12:00:00 AM	1	1/7/2018, 12:00:00 AM	Sunday	SUN	1	7	7	JAN-2018	
8	20180108	1/8/2018, 12:00:00 AM	1	1/8/2018, 12:00:00 AM	Monday	MON	2	8	8	JAN-2018	
9	20180109	1/9/2018, 12:00:00 AM	1	1/9/2018, 12:00:00 AM	Tuesday	TUE	3	9	9	JAN-2018	
10	20180110	1/10/2018, 12:00:00 AM	1	1/10/2018, 12:00:00 AM	Wednesday	WED	4	10	10	JAN-2018	
11	20180111	1/11/2018, 12:00:00 AM	1	1/11/2018, 12:00:00 AM	Thursday	THU	5	11	11	JAN-2018	
12	20180112	1/12/2018, 12:00:00 AM	1	1/12/2018, 12:00:00 AM	Friday	FRI	6	12	12	JAN-2018	
13	20180113	1/13/2018, 12:00:00 AM	1	1/13/2018, 12:00:00 AM	Saturday	SAT	7	13	13	JAN-2018	
14	20180114	1/14/2018, 12:00:00 AM	1	1/14/2018, 12:00:00 AM	Sunday	SUN	1	14	14	JAN-2018	

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

The screenshot shows the Oracle SQL Developer interface. The left pane displays the database schema with the following tables: ADMIN, CUSTOMER_INFO, DATE_INFO, DBTOOLSSEXEUTION_HISTORY, DELIVERY_INFO, DIM_CUSTOMER, DIM_DATE, DIM_PRODUCT, EMPLOYEE, FACT_SALES, PRODUCT_INFO, PROMOTION_DISCOUNT_INFO, SALES_ORDER, STORE, STORE_INFO, and SUPPLIER_INFO. The main pane shows the query result for the query "SELECT * FROM Dim_Date". The query is executed, and the results are displayed in a table with 11 columns: DATEKEY, DATE, DAY_TIME_SPAN, DAY_END_DATE, WEEK_DAY_FULL, WEEK_DAY_SHORT, DAY_NUM_OF_WEEK, DAY_NUM_OF_MONTH, DAY_NUM_OF_YEAR, MONTH_ID, and MONTH_NAME. The results show data for the month of January 2016, starting from 1/1/2016 and ending on 1/13/2016.

DATEKEY	DATE	DAY_TIME_SPAN	DAY_END_DATE	WEEK_DAY_FULL	WEEK_DAY_SHORT	DAY_NUM_OF_WEEK	DAY_NUM_OF_MONTH	DAY_NUM_OF_YEAR	MONTH_ID	MONTH_NAME
20160101	1/1/2016 12:00:00 AM	1	1/1/2016 12:00:00 AM	Friday	FRI	6	1	1	JAN-2016	
20160102	1/2/2016 12:00:00 AM	1	1/2/2016 12:00:00 AM	Saturday	SAT	7	2	2	JAN-2016	
20160103	1/3/2016 12:00:00 AM	1	1/3/2016 12:00:00 AM	Sunday	SUN	1	3	3	JAN-2016	
20160104	1/4/2016 12:00:00 AM	1	1/4/2016 12:00:00 AM	Monday	MON	2	4	4	JAN-2016	
20160105	1/5/2016 12:00:00 AM	1	1/5/2016 12:00:00 AM	Tuesday	TUE	3	5	5	JAN-2016	
20160106	1/6/2016 12:00:00 AM	1	1/6/2016 12:00:00 AM	Wednesday	WED	4	6	6	JAN-2016	
20160107	1/7/2016 12:00:00 AM	1	1/7/2016 12:00:00 AM	Thursday	THU	5	7	7	JAN-2016	
20160108	1/8/2016 12:00:00 AM	1	1/8/2016 12:00:00 AM	Friday	FRI	6	8	8	JAN-2016	
20160109	1/9/2016 12:00:00 AM	1	1/9/2016 12:00:00 AM	Saturday	SAT	7	9	9	JAN-2016	
20160110	1/10/2016 12:00:00 AM	1	1/10/2016 12:00:00 AM	Sunday	SUN	1	10	10	JAN-2016	
20160111	1/11/2016 12:00:00 AM	1	1/11/2016 12:00:00 AM	Monday	MON	2	11	11	JAN-2016	
20160113	1/13/2016 12:00:00 AM	1	1/13/2016 12:00:00 AM	Tuesday	TUE	3	12	12	JAN-2016	

Appendix D: Screenshot of “Select * from Dim_Product”

The screenshot shows the Oracle SQL Developer interface. The left pane displays the database schema with the following tables: ADMIN, CUSTOMER_INFO, DATE_INFO, DBTOOLSSEXEUTION_HISTORY, DELIVERY_INFO, DIM_CUSTOMER, DIM_DATE, DIM_PRODUCT, EMPLOYEE, FACT_SALES, PRODUCT_INFO, PROMOTION_DISCOUNT_INFO, SALES_ORDER, STORE, STORE_INFO, and SUPPLIER_INFO. The main pane shows the query result for the query "Select * from Dim_Product". The query is executed, and the results are displayed in a table with 10 columns: PRODUCTKEY, PRODUCTNAME, CATEGORY, SUBCATEGORY, BRAND, ISCURRENT, PRODUCTID, SCD_START, SCD_END, and VERSIONING. The results show data for products such as Cinnamon Bread, Milk, Chocolate Chip Cookies, Eggs, and Rotini.

PRODUCTKEY	PRODUCTNAME	CATEGORY	SUBCATEGORY	BRAND	ISCURRENT	PRODUCTID	SCD_START	SCD_END	VERSIONING
1	Cinnamon Bread	Wheat	Bread	Nothing Breader	Y	1	1/1/2024 12:00:00 AM	12/31/2099 12:00:00	1
2	Milk	Dairy	Liquid	Buffalo Farms	Y	2	2/1/2024 12:00:00 AM	12/31/2099 12:00:00	1
3	Chocolate Chip Cookies	Candy	Cookies	Nothing Breader	Y	3	3/1/2024 12:00:00 AM	12/31/2099 12:00:00	1
4	Eggs	Dairy	Solid	Rochester Farms	Y	4	4/1/2024 12:00:00 AM	12/31/2099 12:00:00	1
5	Rotini	Wheat	Pasta	Buffalo Farms	Y	5	5/1/2024 12:00:00 AM	12/31/2099 12:00:00	1

Appendix E: Screenshot of test pipeline.

The screenshot shows the Hop ETL tool interface. The top bar indicates the project is 'default' and the environment is 'e_e_x'. The pipeline is named 'testfile' and contains two transforms: 'LoadDimProduct' and 'LoadDimCustomer', with the current view showing 'LOAD_FACT_SALES_STAGING'. The pipeline diagram shows a 'Table input' transform connected to a 'Text file output' transform. The 'Metrics' tab is selected, showing the following data:

#	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

The screenshot shows the Hop ETL tool interface. The top bar indicates the project is 'default' and the environment is 'e_e_x'. The pipeline is named 'testfile' and contains two transforms: 'LoadDimProduct' and 'LoadDimCustomer', with the current view showing 'LOAD_FACT_SALES_STAGING'. The pipeline diagram shows a 'Microsoft Excel input' transform connected to a 'Dimension lookup/update' transform. The 'Metrics' tab is selected, showing the following data:

#	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.

The screenshot shows the Oracle SQL Developer interface. The top bar includes the Oracle logo, 'Database Actions | SQL', a search bar, and a user profile 'ADMIN'. The left sidebar has a 'Navigator' tab with a tree view showing database objects: DBTOOLSSEXEUTION_HISTORY, DIM_CUSTOMER, DIM_DATE, DIM_PRODUCT, and FACT_SALES. The main workspace shows a SQL script: 'Select * from DIM_PRODUCT'. Below the script, the 'Query Result' tab is active, displaying a table with 10 columns: PRODUCTKEY, PRODUCTNAME, CATEGORY, SUBCATEGORY, BRAND, ISCURRENT, PRODUCTID, SCD_START, SCD_END, and VERSIONING. The table contains 7 rows of data.

	PRODUCTKEY	PRODUCTNAME	CATEGORY	SUBCATEGORY	BRAND	ISCURRENT	PRODUCTID	SCD_START	SCD_END	VERSIONING
1	1	Cinnamon Bread Loaf	Wheat	Bread	Nothing Breadr	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 Breadr	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

The screenshot shows the Hop data pipeline interface. The top bar includes the Hop logo, 'Project: default', and 'Environment:'. The main workspace displays a pipeline with two tasks: 'Microsoft Excel input' and 'Dimension lookup/update', connected by a data flow arrow. The bottom panel shows a 'Metrics' table with columns for Transform Name, Copy, Input, Read, Written, Output, Updated, Rejected, Errors, Buffers Input, Buffers Output, Duration, Speed, Status, and a final empty column.

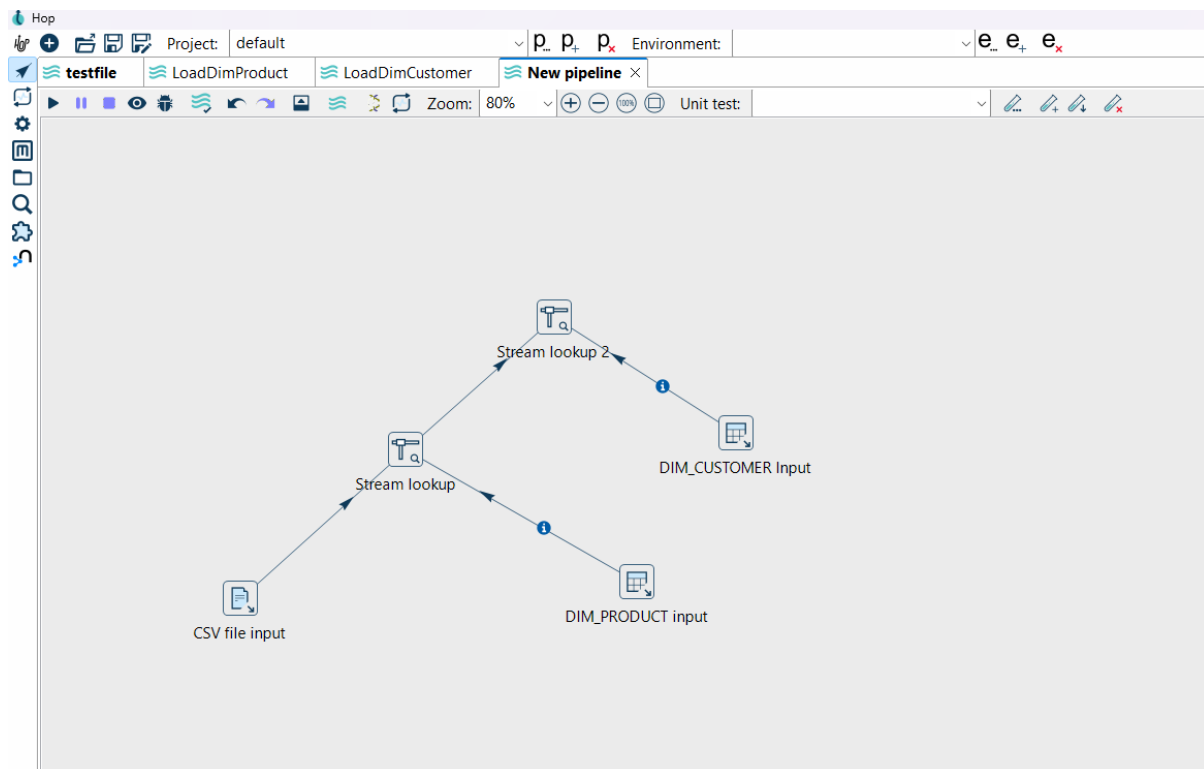
#	Transform Name	Copy	Input	Read	Written	Output	Updated	Rejected	Errors	Buffers Input	Buffers Output	Duration	Speed	Status	
1	Microsoft Excel input	0	5	0	5	0	0	0	0	0	0	0.002"	96	Finished	
2	Dimension lookup/update	0	5	5	5	4	2	0	0	0	0	0.740"	6	Finished	

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

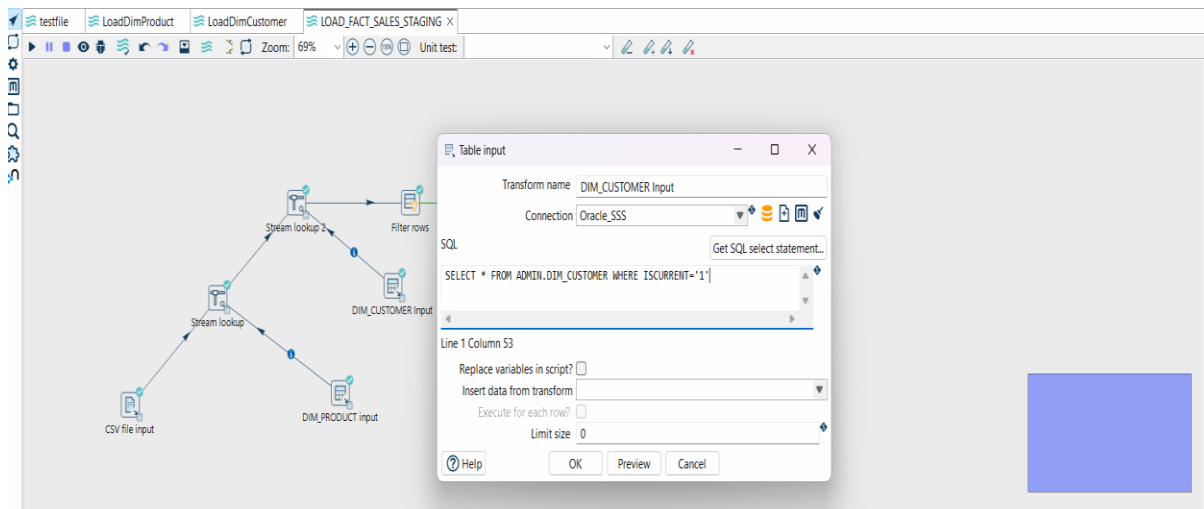
The screenshot shows the Oracle SQL Developer interface. The top bar indicates the user is 'ADMIN' and the database is 'Search Database (Ctrl+L)'. The left sidebar shows the 'Navigator' with a tree view containing 'ADMIN', 'Tables', and a search bar. Below this, a list of tables is shown: 'DETOOLS\$EXECUTION_HISTORY', 'DIM_CUSTOMER', 'DIM_DATE', 'DIM_PRODUCT', and 'FACT_SALES'. The main window displays the query 'Select * from DIM_CUSTOMER' in the 'SQL' tab. The 'Query Result' tab is active, showing a table with 13 columns: 'CUSTOMERKEY', 'CNAME', 'BIRTHDAY', 'ADDRESS', 'CITY', 'STATEPROV', 'ZIP', 'ISCURRENT', 'CUSTID', 'SCD_START', 'SCD_END', and 'VERSIONING'. The table contains 7 rows of data. The execution time is 0.039 seconds.

	CUSTOMERKEY	CNAME	BIRTHDAY	ADDRESS	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

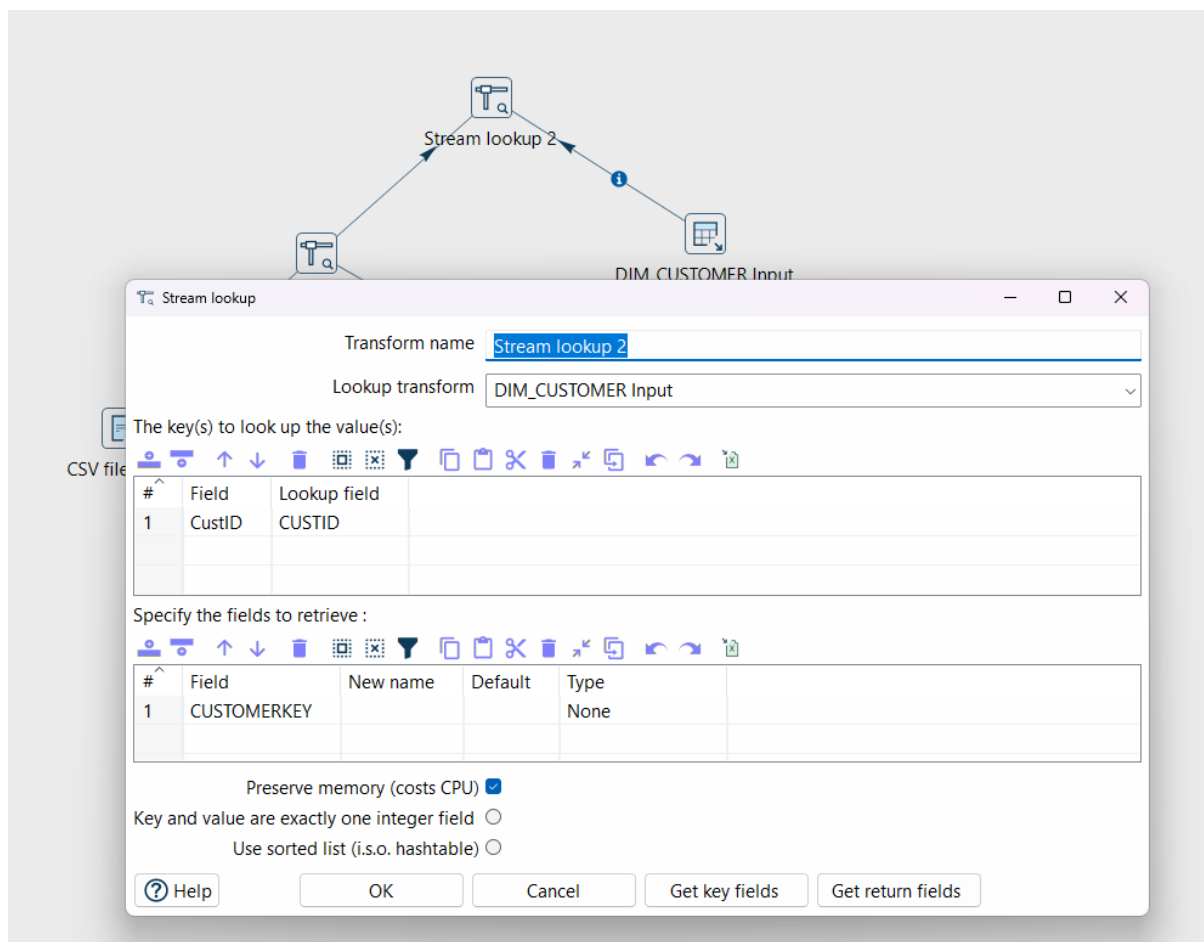
Appendix J: Screenshot of flow of pipeline



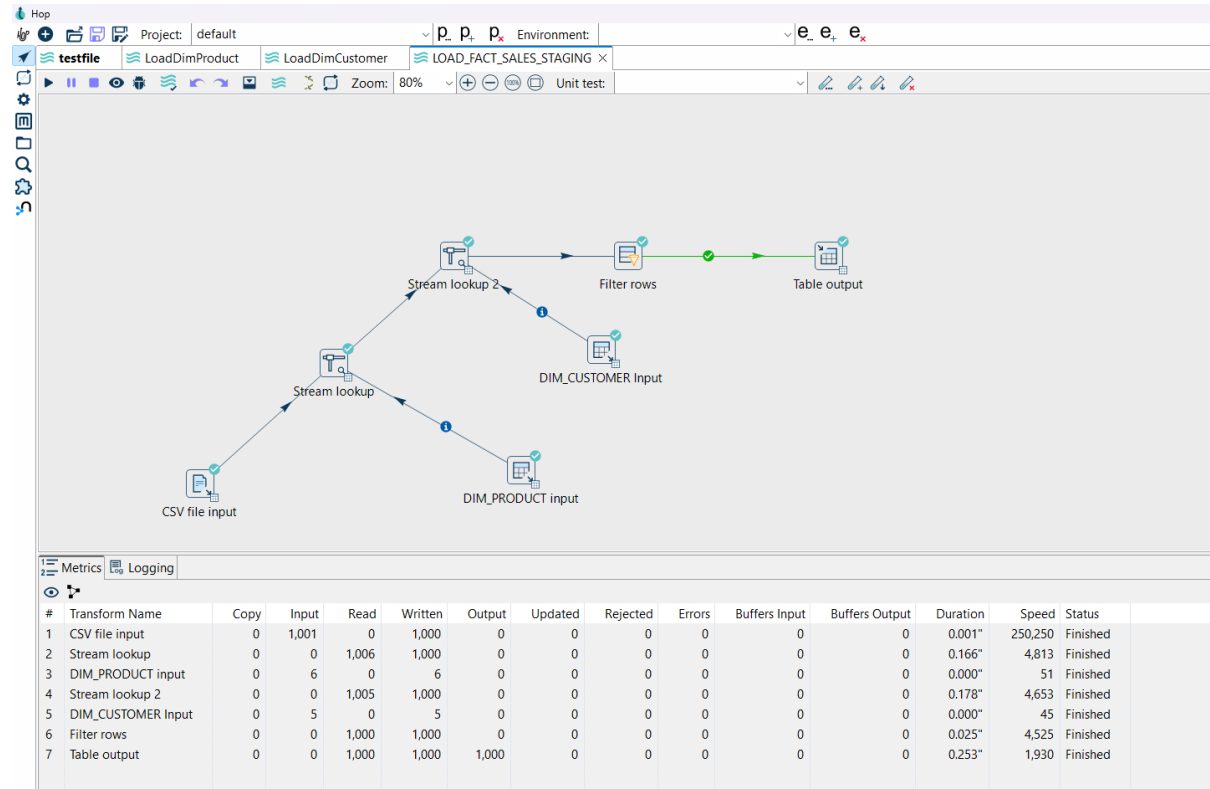
Appendix K: Screenshot of DIM_Customer_input



Appendix L: Screenshot of Stream lookup 2



Appendix M: Screenshot of complete pipeline.



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

The screenshot shows the Oracle SQL Developer interface with the query result for "SELECT * FROM FACT_SALES". The query was executed successfully, and the results are displayed in a table.

	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