Assignment A1 (Data Mining & Warehousing)

Problem Definition:
For an organization of your choice, choose a set of business processes. Design star / snowflake schemas for analyzing these processes. Create a fact constellation schema by combining them. Extract data from different data sources, apply suitable transformations and load into destination tables using an ETL tool. For Example: Business Origination: Sales, Order, Marketing Process.

Requirements:
●Windows 10 ·
●Open source ETL tool for Linux/Windows - Pentaho

Learning Objectives:

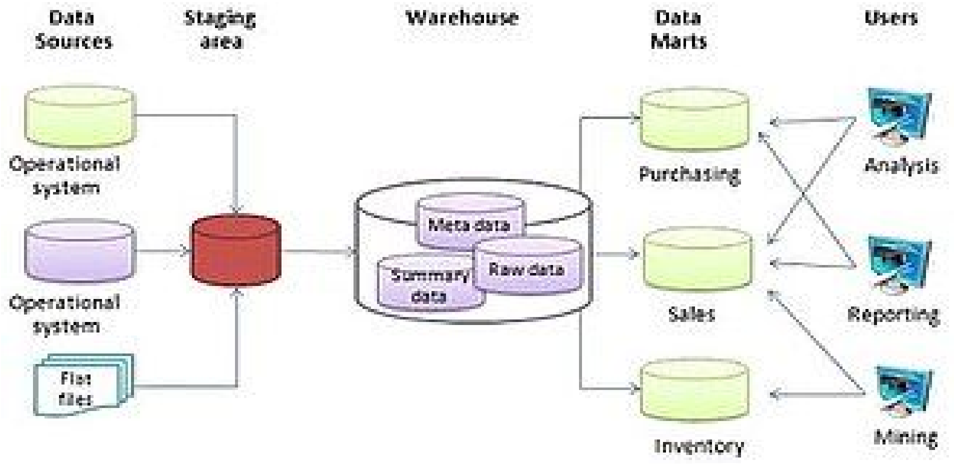
●To understand the concept of Data cube and multidimensional model for data warehouse.
●To understand different preprocessing techniques.
●To study ETL tool Pentaho.

Learning Outcomes:

●Understood the concept of data cube and multidimensional model for data warehouse.
●Applied preprocessing technique in order to replace missing data in the dataset.
●Used pentaho ETL tool to extract data from sales data set, apply suitable transformations and load into destination tables.

Theory:

- **Data Warehouse**
  A data warehouse is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence. Business analysts, data engineers, data scientists, and decision makers access the data through business intelligence (BI) tools, SQL clients, and other analytics applications.A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count,or sum(sales amount). A data cube provides a multidimensional view of data and allows the pre-computation and fast access of summarized data. A multidimensional model views data in the form of a data-cube. A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.



Overview

- **Data Warehousing**

  Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

  Data Warehousing allows for the following:
  - Tuning Strategies
  - Customer analysis
  - Product analysis

  In order to derive value from your data, you need not only have it in one place and using a single, canonical access language and interface, but you must also have a means to manage metadata, handle governance issues, and scale as your data grows. These are among the many challenges that data warehousing tools solve.

  Functions of Data Warehousing tools:
  - Data Extraction − Involves gathering data from multiple heterogeneous sources.

  - Data Cleaning − Involves finding and correcting the errors in data.

  - Data Transformation − Involves converting the data from legacy format to warehouse format.

  - Data Loading − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.

    - Refreshing − Involves updating from data sources to warehouse

- **ETL Data Warehousing Tool**
  ETL stands for "Extract, Transform, and Load" and consists of the tools and processes used for pulling data from one store, transforming it for placement, and finally, loading it into another (often aggregate) store. Just as with data warehouses, ETL tools have progressed over time from self-administered to cloud-native offerings.

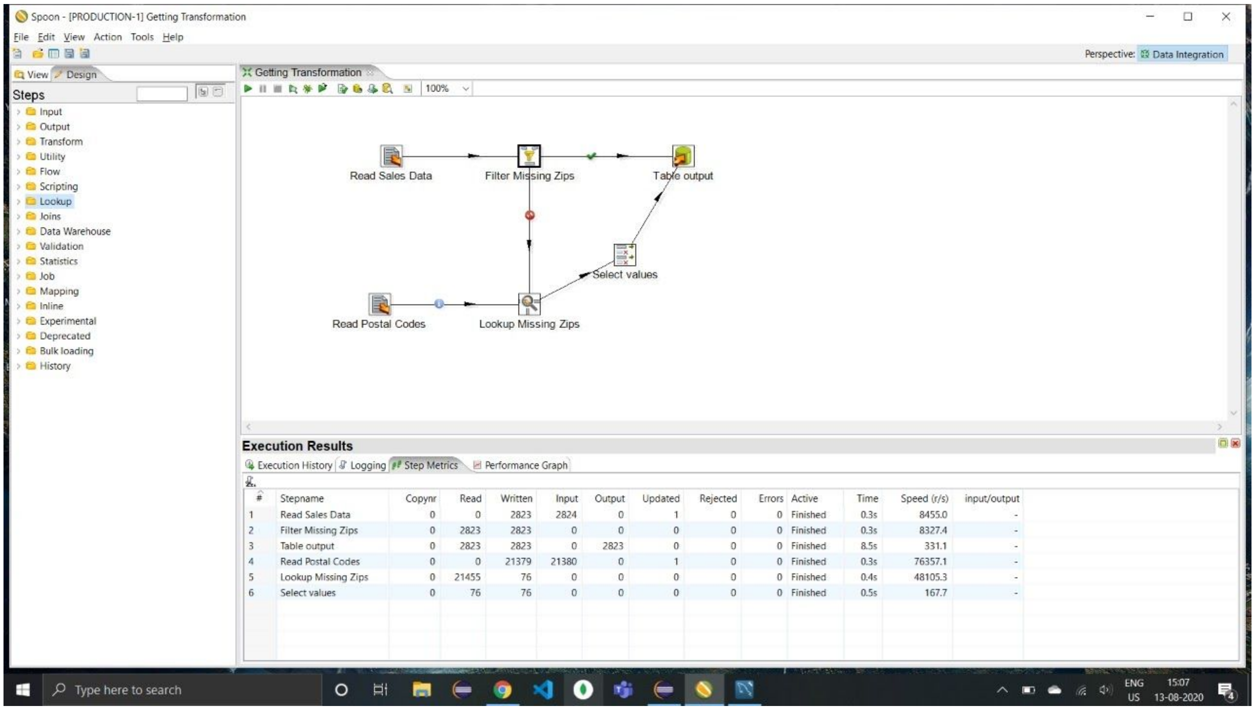
- **Star Schema**
  Star schema is a mature modeling approach widely adopted by relational data warehouses. It requires modelers to classify their model tables as either dimension or fact. Dimension tables describe business entities—the things you model. Entities can include products, people, places, and concepts including time itself. The most consistent table you'll find in a star schema is a date dimension table. A dimension table contains a key column (or columns) that acts as a unique identifier, and descriptive columns. Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc. A fact table contains dimension key columns that relate to dimension tables, and numeric measure columns. The dimension key columns determine the dimensionality of a fact table, while the dimension key values determine the granularity of a fact table.

- **Snowflake Schema**
  It is a logical arrangement of tables in a multidimensional database such that the ER diagram resembles a snowflake shape. A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. The dimension tables are normalized which splits data into additional tables. The main benefit of the snowflake schema is that it uses smaller disk space. Easier to implement a dimension is added to the Schema
  Due to multiple tables query performance is reduced. The primary challenge that you will face while using the Snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

Star Schema	**Snowflake Schema**
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
In a star schema, only single join creates the relationship between the fact table and any dimension tables	A snowflake schema requires many joins to fetch the data.
Simple DB Design.	Very Complex DB Design.
Denormalized Data structure and query also run faster.	Normalized Data Structure.
High level of Data redundancy	Very low-level data redundancy

OPERATIONS PERFORMED:

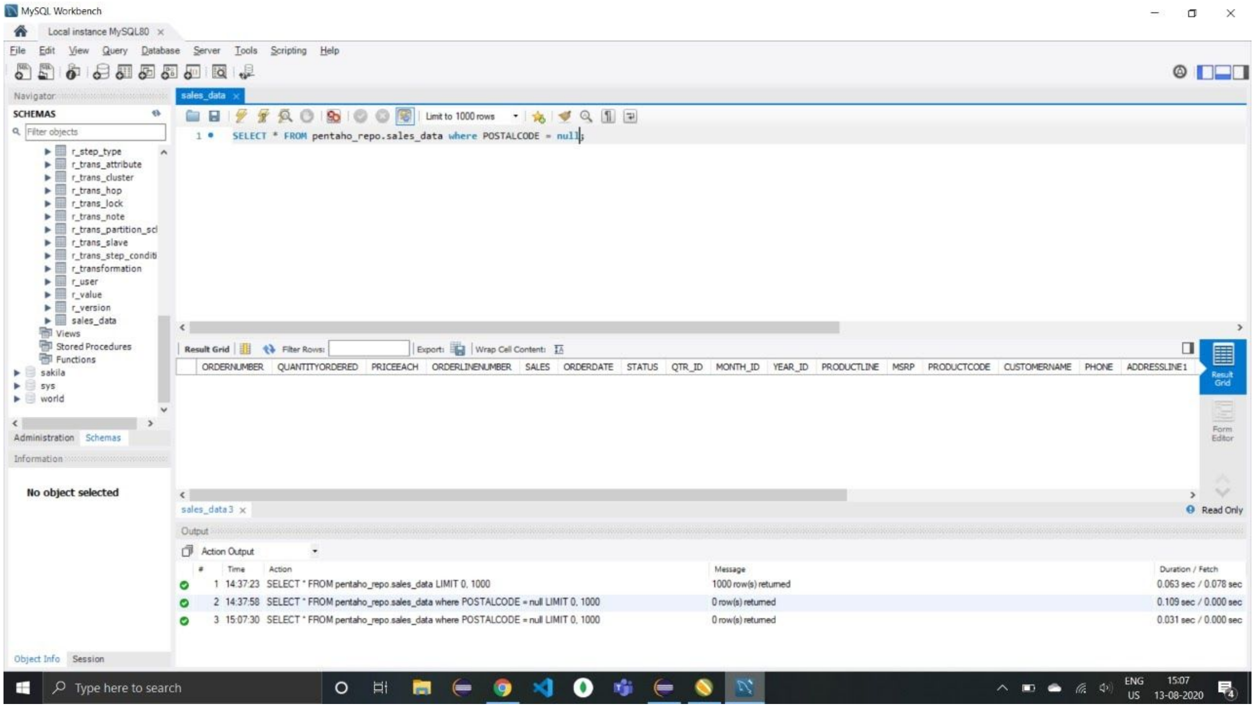
1. Retrieve sales data from sample sales data csv file (Extraction).
2. Filter records with missing postal codes (POSTALCODE = null).
3. Load the data into MySQL database.
4. Perform lookup operation for missing zip codes.
5. Resolve missing zip codes.
6. Write the resolved data into the database.



Execution Results

Execution History | Logging | Step Metrics | Performance Graph

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Read Sales Data	0	0	2823	2824	0	1	0	0	Finished	0.3s	8455.0	-
2	Filter Missing Zips	0	2823	2823	0	0	0	0	0	Finished	0.3s	8327.4	-
3	Table output	0	2823	2823	0	2823	0	0	0	Finished	8.5s	331.1	-
4	Read Postal Codes	0	0	21379	21380	0	1	0	0	Finished	0.3s	76357.1	-
5	Lookup Missing Zips	0	21455	76	0	0	0	0	0	Finished	0.4s	48105.3	-
6	Select values	0	76	76	0	0	0	0	0	Finished	0.5s	167.7	-



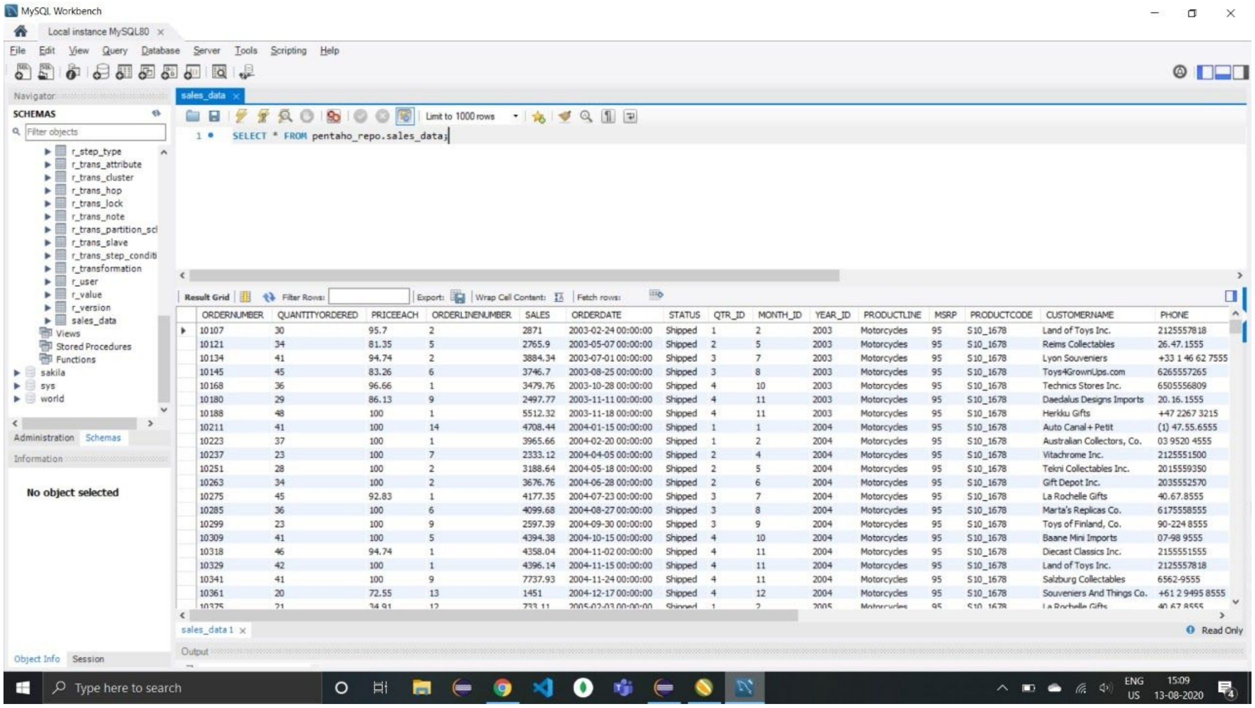
```
SELECT * FROM pentaho_repo.sales_data where POSTALCODE = null;
```

ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODUCTLINE	MSRP	PRODUCTCODE	CUSTOMERNAME	PHONE	ADDRESSLINE1

Output

Action Output

```
#  Time      Action                                                                    Message              Duration / Fetch
1  14:37:23  SELECT * FROM pentaho_repo.sales_data LIMIT 0, 1000                         1000 row(s) returned  0.063 sec / 0.078 sec
2  14:37:58  SELECT * FROM pentaho_repo.sales_data where POSTALCODE = null LIMIT 0, 1000  0 row(s) returned     0.109 sec / 0.000 sec
3  15:07:30  SELECT * FROM pentaho_repo.sales_data where POSTALCODE = null LIMIT 0, 1000  0 row(s) returned     0.031 sec / 0.000 sec
```



```
SELECT * FROM pentaho_repo.sales_data;
```

ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODUCTLINE	MSRP	PRODUCTCODE	CUSTOMERNAME	PHONE
10107	30	95.7	2	2871	2003-02-24 00:00:00	Shipped	1	2	2003	Motorcycles	95	S10_1678	Land of Toys Inc.	2125557818
10121	34	81.35	5	2765.9	2003-05-07 00:00:00	Shipped	2	5	2003	Motorcycles	95	S10_1678	Reims Collectables	26.47.1555
10134	41	94.74	2	3884.34	2003-07-01 00:00:00	Shipped	3	7	2003	Motorcycles	95	S10_1678	Lyon Souveniers	+33 1 46 62 7555
10145	45	83.26	6	3746.7	2003-08-25 00:00:00	Shipped	3	8	2003	Motorcycles	95	S10_1678	Toys4GrownUps.com	6265557265
10168	36	96.66	1	3479.76	2003-10-28 00:00:00	Shipped	4	10	2003	Motorcycles	95	S10_1678	Technics Stores Inc.	6505556809
10180	29	86.13	9	2497.77	2003-11-11 00:00:00	Shipped	4	11	2003	Motorcycles	95	S10_1678	Daedalus Designs Imports	20.16.1555
10188	48	100	1	5512.32	2003-11-18 00:00:00	Shipped	4	11	2003	Motorcycles	95	S10_1678	Herkku Gifts	+47 2267 3215
10211	41	100	14	4708.44	2004-01-15 00:00:00	Shipped	1	1	2004	Motorcycles	95	S10_1678	Auto Canal+Petit	(1) 47.55.6555
10223	37	100	1	3965.66	2004-02-20 00:00:00	Shipped	1	2	2004	Motorcycles	95	S10_1678	Australian Collectors, Co.	03 9520 4555
10237	23	100	7	2333.12	2004-04-05 00:00:00	Shipped	2	4	2004	Motorcycles	95	S10_1678	Vitachrome Inc.	2125551500
10251	28	100	2	3188.64	2004-05-18 00:00:00	Shipped	2	5	2004	Motorcycles	95	S10_1678	Tekni Collectables Inc.	2015559350
10263	34	100	2	3676.76	2004-06-28 00:00:00	Shipped	2	6	2004	Motorcycles	95	S10_1678	Gift Depot Inc.	2035552570
10275	45	92.83	1	4177.35	2004-07-23 00:00:00	Shipped	3	7	2004	Motorcycles	95	S10_1678	La Rochelle Gifts	40.67.8555
10285	36	100	6	4099.68	2004-08-27 00:00:00	Shipped	3	8	2004	Motorcycles	95	S10_1678	Marta's Replicas Co.	6175558555
10299	23	100	9	2597.39	2004-09-30 00:00:00	Shipped	3	9	2004	Motorcycles	95	S10_1678	Toys of Finland, Co.	90-224 8555
10309	41	100	5	4394.38	2004-10-15 00:00:00	Shipped	4	10	2004	Motorcycles	95	S10_1678	Baane Mini Imports	07-98 9555
10318	46	94.74	1	4358.04	2004-11-02 00:00:00	Shipped	4	11	2004	Motorcycles	95	S10_1678	Diecast Classics Inc.	2155551555
10329	42	100	1	4396.14	2004-11-15 00:00:00	Shipped	4	11	2004	Motorcycles	95	S10_1678	Land of Toys Inc.	2125557818
10341	41	100	9	7737.93	2004-11-24 00:00:00	Shipped	4	11	2004	Motorcycles	95	S10_1678	Salzburg Collectables	6562-9555
10361	20	72.55	13	1451	2004-12-17 00:00:00	Shipped	4	12	2004	Motorcycles	95	S10_1678	Souveniers And Things Co.	+61 2 9495 8555
10375	21	34.91	12	733.11	2005-02-03 00:00:00	Shipped	1	2	2005	Motorcycles	95	S10_1678	La Rochelle Gifts	40.67.8555

Conclusion:

Used pentaho ETL tool to extract data from sales data set, applied preprocessing technique in order to replace missing zip codes and loaded the data into destination tables