DMW - Assignment No. 1 **Roll no : 41239**

Date : 20/8/20

Title:

For an organization of your choice, choose a set of business processes. Design star / snow flake 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.

Problem Statement:

Design a basic ETL model using Pentaho Data Integration tool.

Objective:

Understand the basics of Star/Snowflake/fact Constellation schema & learn the Pentaho tool for per perform various Operation on in-built or external Datasets.

Outcomes:

1. Students will be able to demonstrate Installation of Pentaho Tool
2. Students will be able to demonstrate different Operator & Datasets in Pentaho
3. Students will be able to demonstrate different Operations on Available data in Pentaho

Hardware Requirement: Any CPU with Pentium Processor or similar, 256 MB RAM or more,1 GB Hard Disk or more

Software Requirements: 32/64 bit Linux/Windows Operating System, Pentaho Tool 4.1

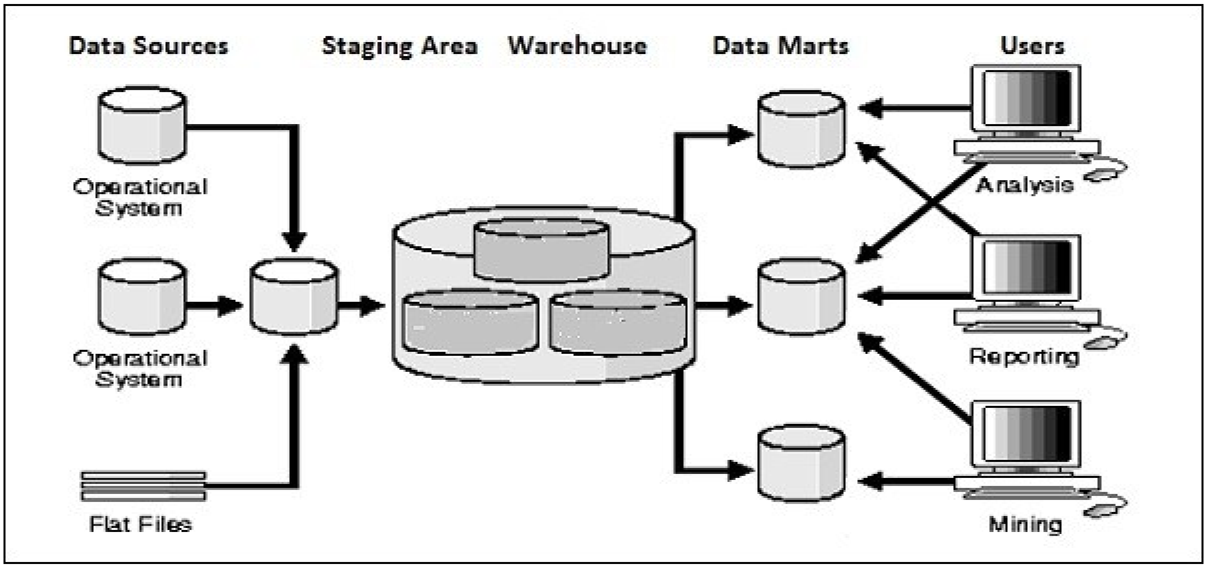
Theory:

What does ETL mean?

ETL stands for Extract, Transform and Load. An ETL tool extracts the data from different RDBMS source systems, transforms the data like applying calculations, concatenate, etc. and then load the data to Data Warehouse system. The data is loaded in the DW system in the form of dimension and fact tables.

Extraction

- A staging area is required during ETL load. There are various reasons why staging area is required.
- The source systems are only available for specific period of time to extract data. This period of time is less than the total data-load time. Therefore, staging area allows you to extract the data from the source system and keeps it in the staging area before the time slot ends.
- Staging area is required when you want to get the data from multiple data sources together or if you want to join two or more systems together. For example, you will not be able to perform a SQL query joining two tables from two physically different databases.
- Data extractions' time slot for different systems vary as per the time zone and operational hours.
- Data extracted from source systems can be used in multiple data warehouse system, Operation Data stores, etc.
- ETL allows you to perform complex transformations and requires extra area to store the data.



Transform

In data transformation, you apply a set of functions on extracted data to load it into the target system. Data, which does not require any transformation is known as direct move or pass through data.

You can apply different transformations on extracted data from the source system. For example, you can perform customized calculations. If you want sum-of-sales revenue and this is not in database, you can apply the **SUM** formula during transformation and load the data.

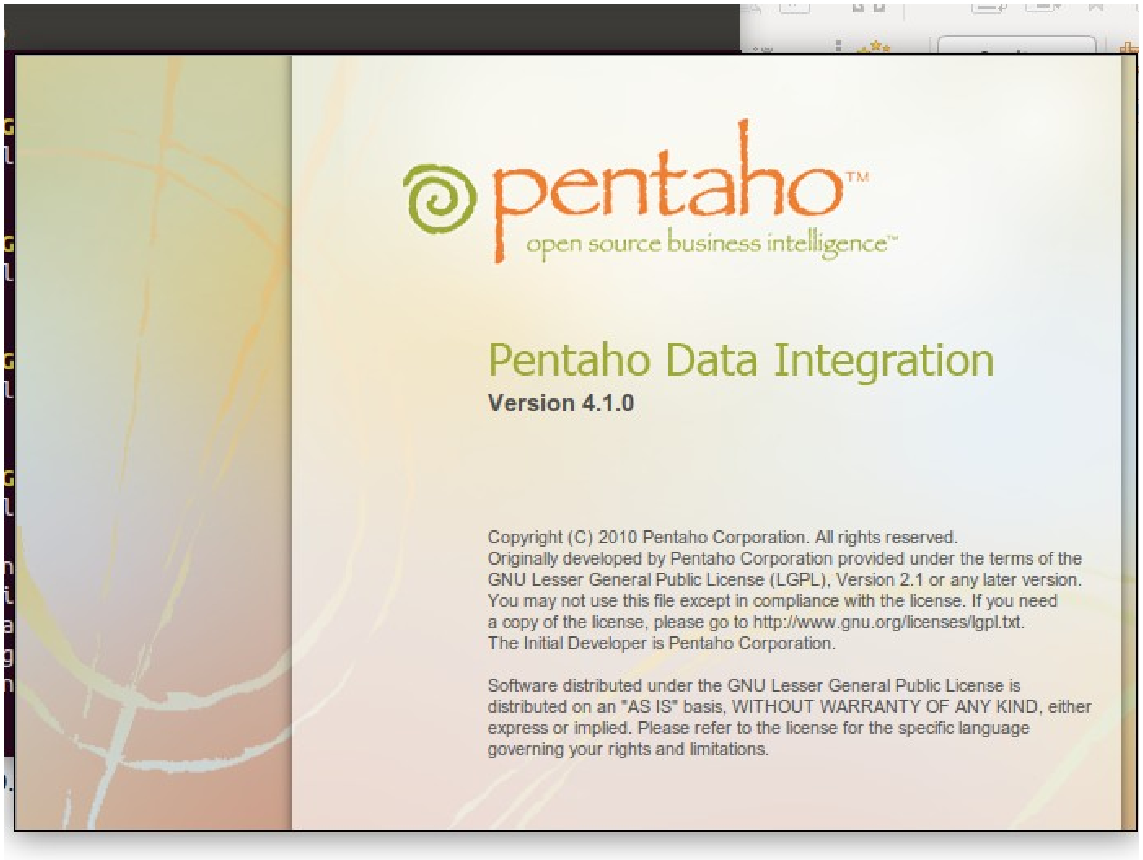
For example, if you have the first name and the last name in a table in different columns, you can use concatenate before loading.

Load

During Load phase, data is loaded into the end-target system and it can be a flat file or a Data Warehouse system.

STEPS FOR INSTALLATION:

1. Create the Pentaho User.
2. Create Linux Directory Structure.
3. Install Java.
4. Install the Web Application Server, if you are installing on your own web application server.
5. Install the Pentaho Repository Host Database.
6. Download and Unpack the Installation Files.
7. Set Environment Variables.
8. Advanced Linux Considerations.



Pentaho Data Integration

Version 4.1.0

Copyright (C) 2010 Pentaho Corporation. All rights reserved.
Originally developed by Pentaho Corporation provided under the terms of the GNU Lesser General Public License (LGPL), Version 2.1 or any later version. You may not use this file except in compliance with the license. If you need a copy of the license, please go to http://www.gnu.org/licenses/lgpl.txt.
The Initial Developer is Pentaho Corporation.

Software distributed under the GNU Lesser General Public License is distributed on an "AS IS" basis, WITHOUT WARRANTY OF ANY KIND, either express or implied. Please refer to the license for the specific language governing your rights and limitations.

Field	Setting
Connection Name:	Sample Data
Connection Type:	MySQL
Host Name	localhost
Database Name	sampledata
Port Number	3306
User Name	Root
Password	password

Step:

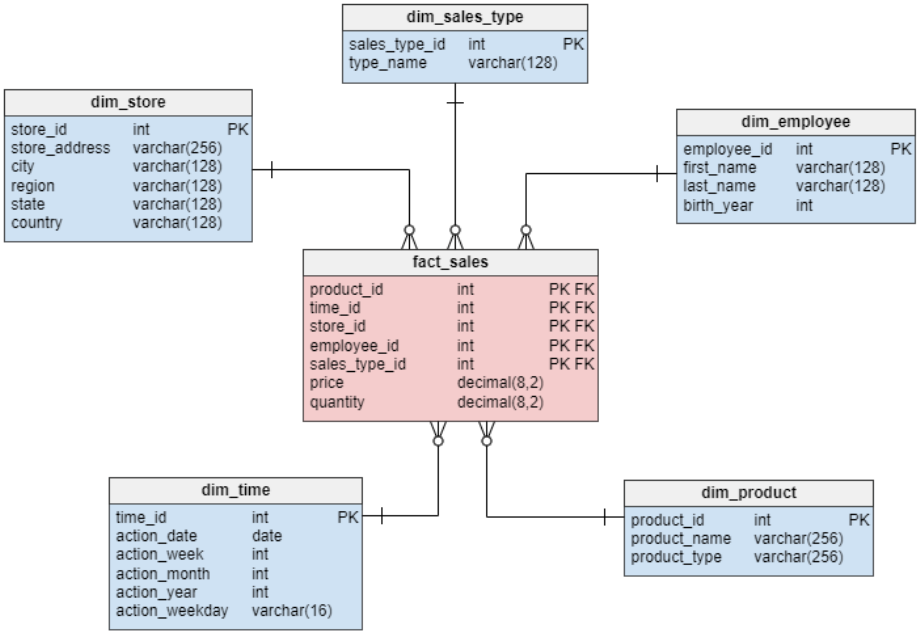
1. Click **OK**, to exit the **Database Connections** window.
2. Type **SALES_DATA** in the **Target Table** text field.
3. Since this table does not exist in the target database, you will need use the software to generate the Data Definition Language (DDL) to create the table and execute it. DDLs are the SQL commands that define the different structures in a database such as CREATE TABLE.
4. In the **Table Output** window, enable the **Truncate Table** property.Click the **SQL** button at the bottom of the **Table output** dialog box to generate the DDL for creating your target table.
5. The **Simple SQL editor** window appears with the SQL statements needed to create the table.
6. Click **Execute** to execute the SQL statement.

Data Warehousing Schemas

1. Star Schema
2. Snowflake Schema
3. Fact Constellation

Star Schema

For example, as you can see in the above-given image that fact table is at the center which contains keys to every dimension table like Deal_ID, Model ID, Date_ID, Product_ID, Branch_ID & other attributes like Units sold and revenue.



Characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are not normalized. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools

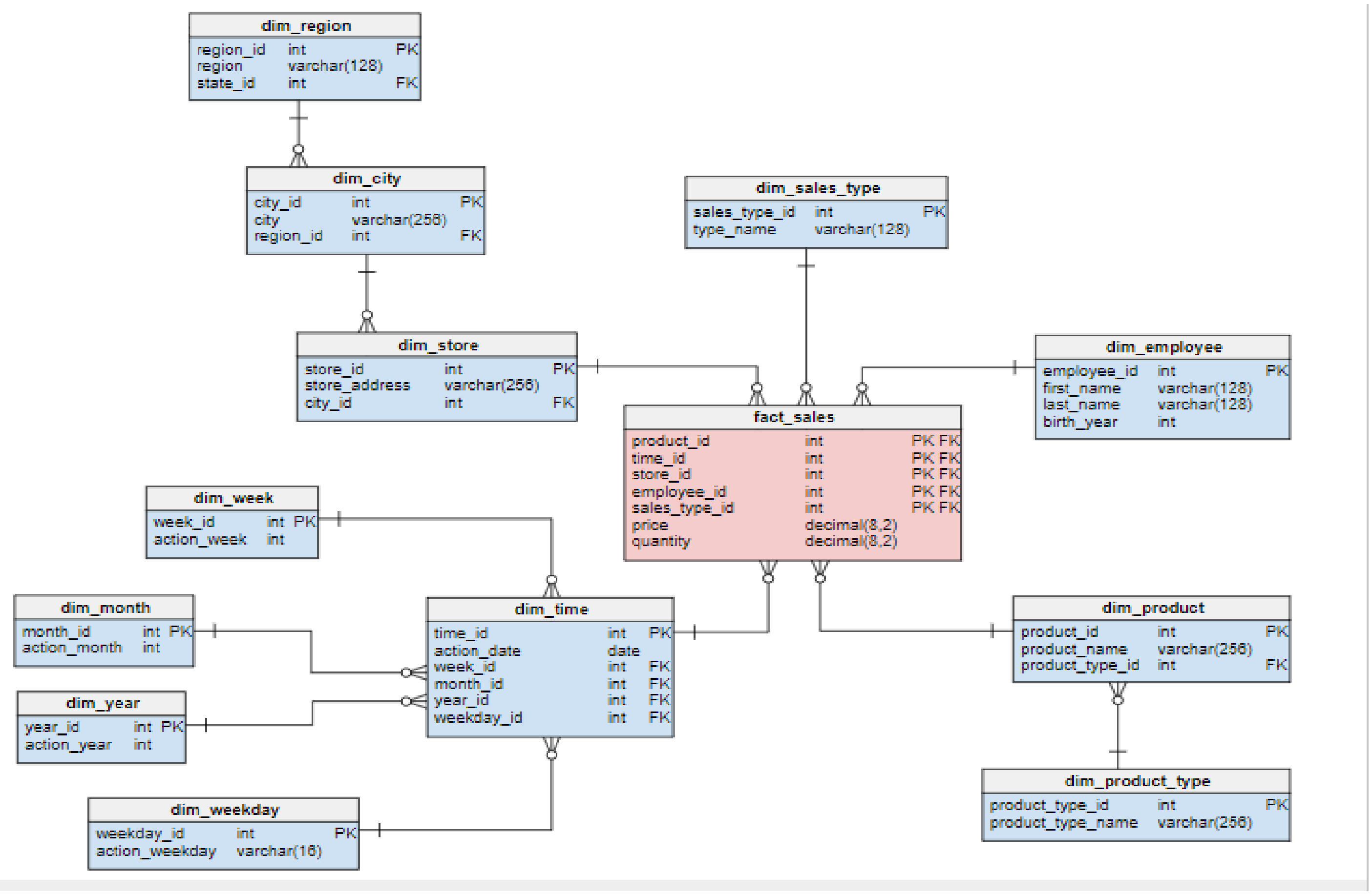
Snowflake Schema

A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. It is called snowflake because its diagram resembles a Snowflake.

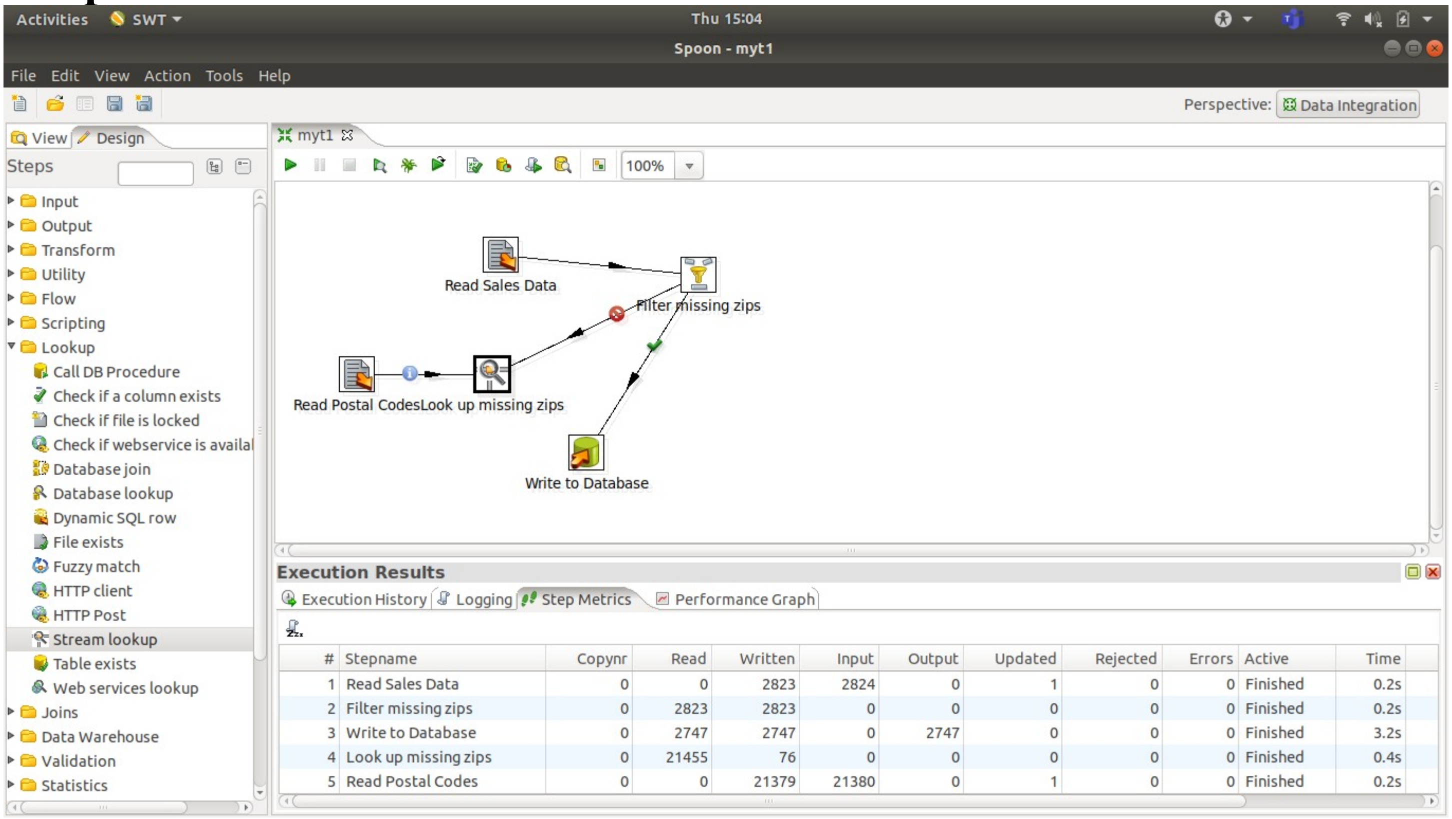
The dimension tables are normalized which splits data into additional tables. In the following example, Country is further normalized into an individual table.

Characteristics of Snowflake Schema:

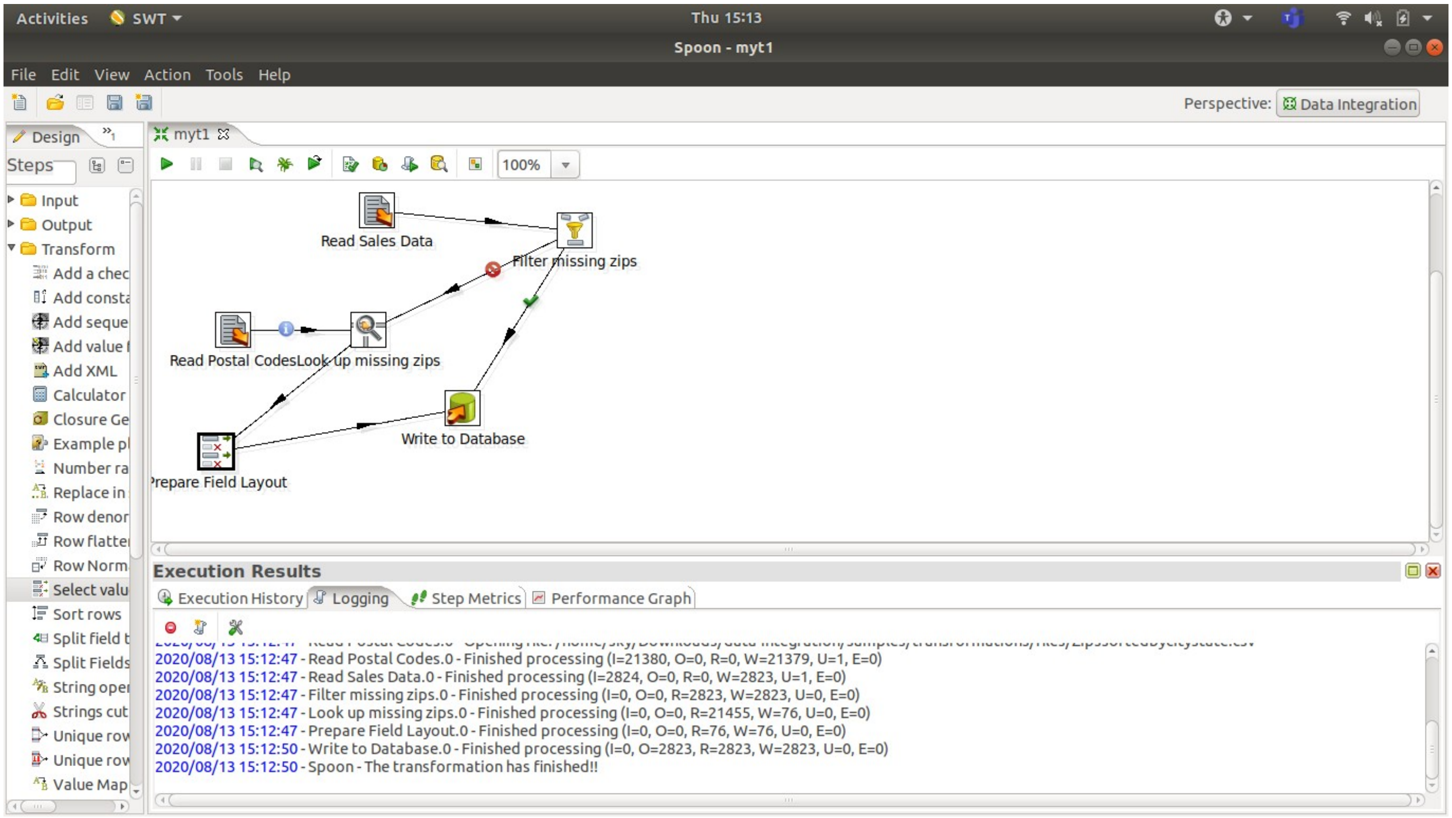
- The main benefit of the snowflake schema it uses smaller disk space.
- Easier to implement a dimension is added to the Schema
- Due to multiple tables query performance is reduced



Output :



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time
1	Read Sales Data	0	0	2823	2824	0	1	0	0	Finished	0.2s
2	Filter missing zips	0	2823	2823	0	0	0	0	0	Finished	0.2s
3	Write to Database	0	2747	2747	0	2747	0	0	0	Finished	3.2s
4	Look up missing zips	0	21455	76	0	0	0	0	0	Finished	0.4s
5	Read Postal Codes	0	0	21379	21380	0	1	0	0	Finished	0.2s



```
2020/08/13 15:12:47 - Read Postal Codes.0 - Opening file: /home/sky/Downloads/data-integration/samples/transformations/files/zipssortedbycitystate.csv
2020/08/13 15:12:47 - Read Postal Codes.0 - Finished processing (I=21380, O=0, R=0, W=21379, U=1, E=0)
2020/08/13 15:12:47 - Read Sales Data.0 - Finished processing (I=2824, O=0, R=0, W=2823, U=1, E=0)
2020/08/13 15:12:47 - Filter missing zips.0 - Finished processing (I=0, O=0, R=2823, W=2823, U=0, E=0)
2020/08/13 15:12:47 - Look up missing zips.0 - Finished processing (I=0, O=0, R=21455, W=76, U=0, E=0)
2020/08/13 15:12:47 - Prepare Field Layout.0 - Finished processing (I=0, O=0, R=76, W=76, U=0, E=0)
2020/08/13 15:12:50 - Write to Database.0 - Finished processing (I=0, O=2823, R=2823, W=2823, U=0, E=0)
2020/08/13 15:12:50 - Spoon - The transformation has finished!!
```

Conclusion:

Hence we are able to study Pentaho Tools us can Perform ETL operations on Sample Data sets and can perform analysis on sample data sets.