DMW-1

Title: Design a multidimensional Data Cube

Problem Statement: For an organization, design star/snowflake schema for analyzing these processes. Create a fact constellation schema by combining them. Extract data from different sources, apply suitable transformations & load into destination tables using ETL.

Objective: - To understand concept of Data cube
- Understand different preprocessing techniques.
- Study ETL tool.

Outcomes: Student will be able to:
- Understand data cube
- Understand preprocessing techniques.
- Study ETL Tool.

S/w & H/w Requirements: Fedora / windows 10, open source ETL Tool.

Theory:

A data warehouse is an integrated & non visible collection of data in support of mangement's decision making process. A data warehouse is usually modelled by a multidimensional data structure, called data cube. A data cube provides a multidimensional view of data & allows the fast access of summarized data.

Data warehouse system use back-end tools & utilities to populate & refresh their data including:

- Data Extraction
- Data cleaning
- Load
- Refresh

The most popular data model for a data warehouse is a multi dimensional model which can exist in the form of a star schema, a snowflake or fact constellation schema.

1. Star Schema:

- Every dimension is represented with only 1-dimension table.
- Dimension table should contain the set of attributes.
- Dimension table is joined to fact table using a foreign key.
- Dimension table are not joined to each other.
- Fact table would contain key & measure.
- Easy to understand & provides optimal disk usage.

2. Snowflake schema:

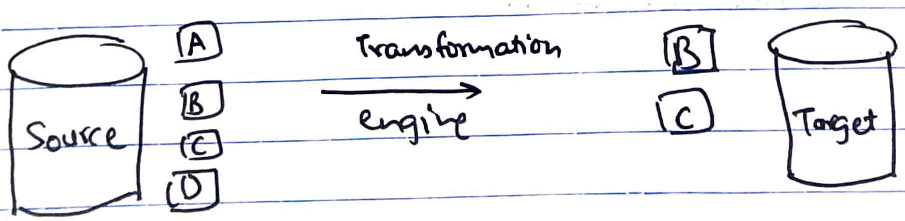
- Uses smaller disk space.
- Easier to implement if a dimension is added.
- Due to multiple tables, query performance is reduced.
- Primary challenge is that you need to perform more maintainance efforts.

3. Fact constellation:

Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars & hence is called galaxy schema.

ETL : Extract Transform Load :

An ETL tool extracts data from different RDB source systems then transforms data, then loads into data warehouse system.



Pentanho:

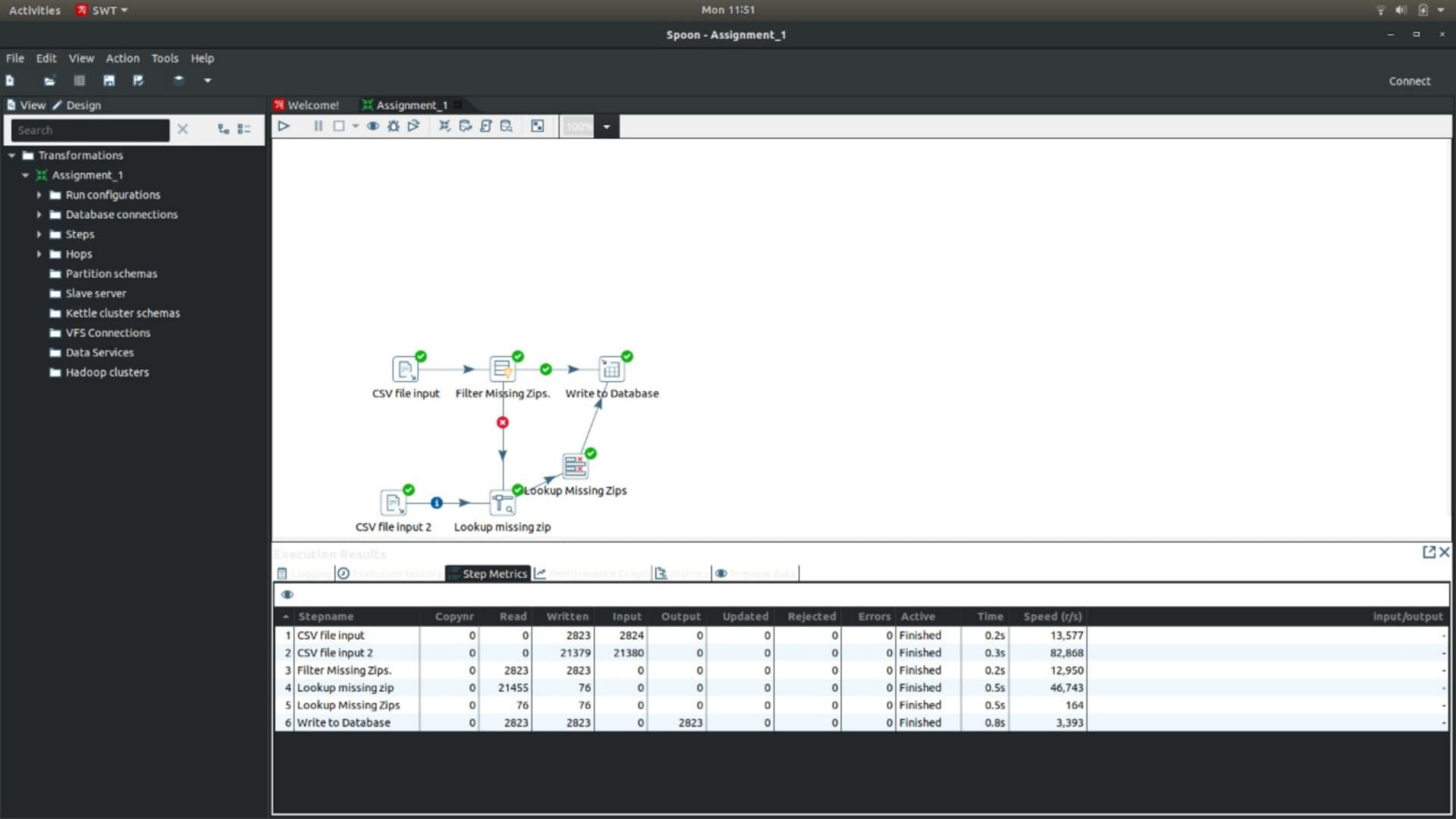
with an intuitive, graphical, drag & drop design experiment & a proven, Scalable, standards based structure. Data integration is increasingly the choice for organization over traditional proprietary ETL or DI Tools. Data Integration delivers powerful ETL capabilities.

Steps for processing using Pentaho:

1. Retrieve data from a flat file after connecting to repository.
2. Use the filter rows transformation, to seperate out those records that have missing postal codes.
3. Take all records existing the previous step where the POSTAL CODE was not NULL & load them into a database table.
4. Next, retrieve data from the lookup file & resolve missing postal codes.
5. Lastly, clean up field layout on lookup stream & run the transformation.

Conclusion:

We learnt to extract data from different Sources, apply suitable transformations & load into destination tables.



	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	CSV file input	0	0	2823	2824	0	0	0	0	Finished	0.2s	13,577	-
2	CSV file input 2	0	0	21379	21380	0	0	0	0	Finished	0.3s	82,868	-
3	Filter Missing Zips.	0	2823	2823	0	0	0	0	0	Finished	0.2s	12,950	-
4	Lookup missing zip	0	21455	76	0	0	0	0	0	Finished	0.5s	46,743	-
5	Lookup Missing Zips	0	76	76	0	0	0	0	0	Finished	0.5s	164	-
6	Write to Database	0	2823	2823	0	2823	0	0	0	Finished	0.8s	3,393	-