DATE:

# Assignment No : DMW-1

* Title : ETL Tool

* Problem Statement :

For an organisation of your choice, choose a set of business process. 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 organization : sales, order, marketing process.

* Objectives : To,

1) learn data modeling concepts and schemas like star, snowflake, fast constellation.
2) apply suitable transforms on data using ETL tool.

* Outcomes : Students will be able to

1) learn and implement data modeling concepts and schemas like star, snowflake, fact constellation.
2) use ETL tool.

* S/W and H/W requirements :

- Java JDK 1.8, MySQL, 5.5 Pentaho pdi-ce-7, MySQL connector - 64-bit OS- Windows 10, Intel core i5 processor, 500 GB HDD.

* Concepts related to theory :

Datawarehousing -

Generalize and consolidate data on multidimensional

space.

OLAP - online Analytic Processing tools, information processing from historical data.

ETL Tool -

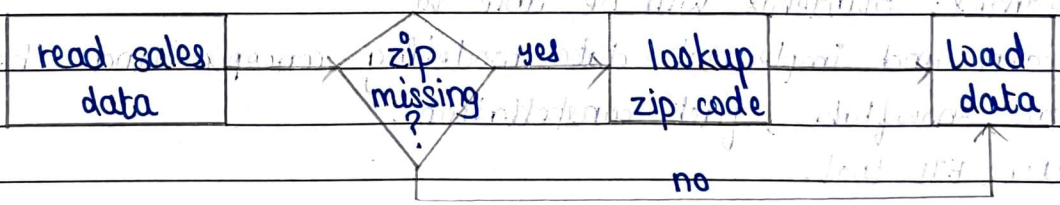
Pentaho Data Integration (PDI) provides access to an extraction, transformation and loading engine that captures night data, cleanses data and stores data using uniform format.

Windows Installation. -

1) Download pentaho pdi-ce-7
2) Install required softwares - JDK 1.8, MySQL server
3) Run spoon.bat file from Data Integration folder.

Assignment steps: ETL using Pentaho on sales data.csv

1) Preprocessing → zip code resolving.




2) Lookup for missing data in zipcodes file

3) Load data to MySQL database → then point to a csv file.

Transformations on data:

a) calculate selling price
b) profit/loss calculation

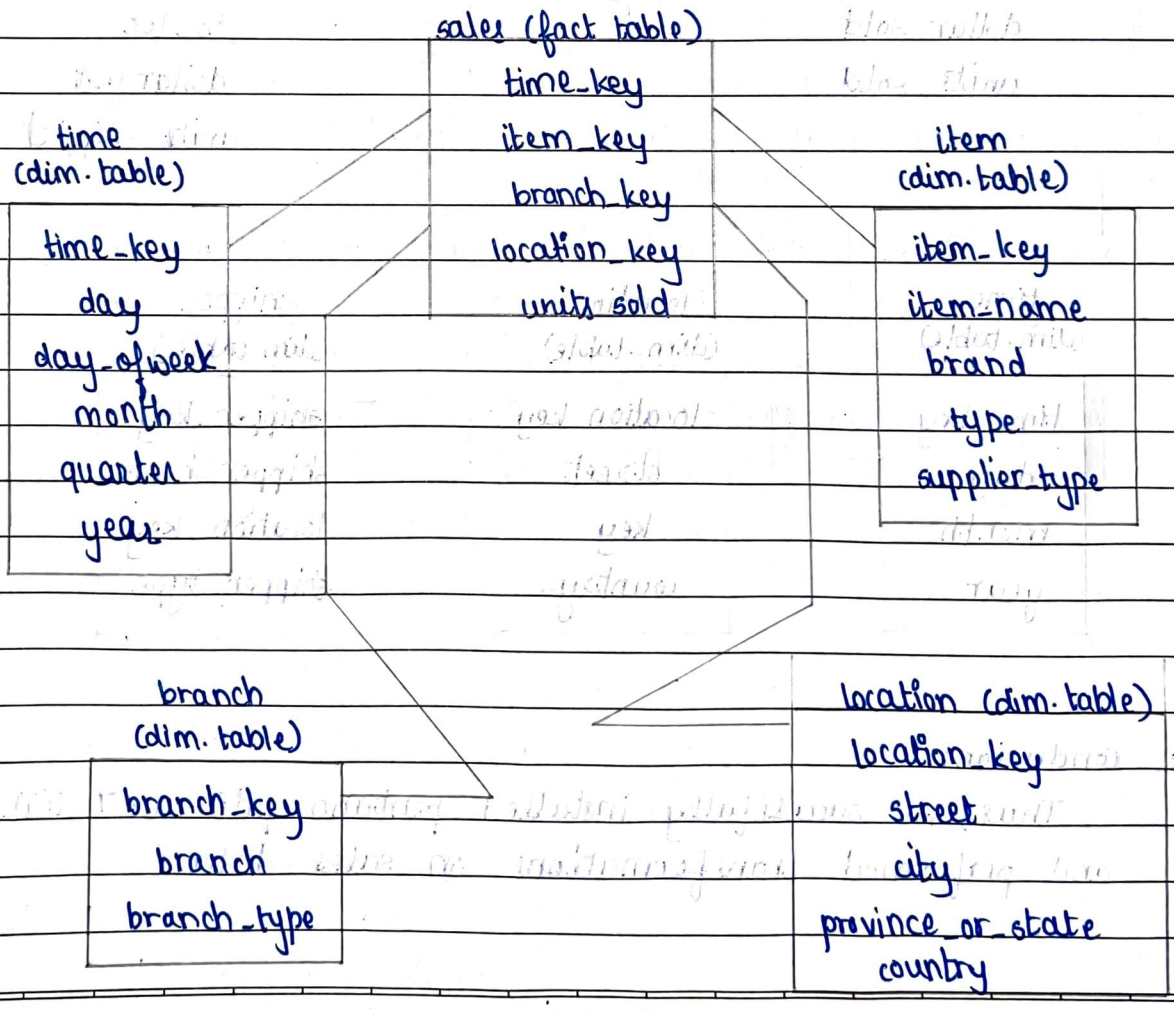
c) concatenation of fields related to product.

Data preprocessing before mining (ETL)

- Data extraction : gathering data.
- Data cleaning : find errors
- Data transformation : convert to different format
- load : summarize views
- Refresh or update.

Modeling data - schemas :

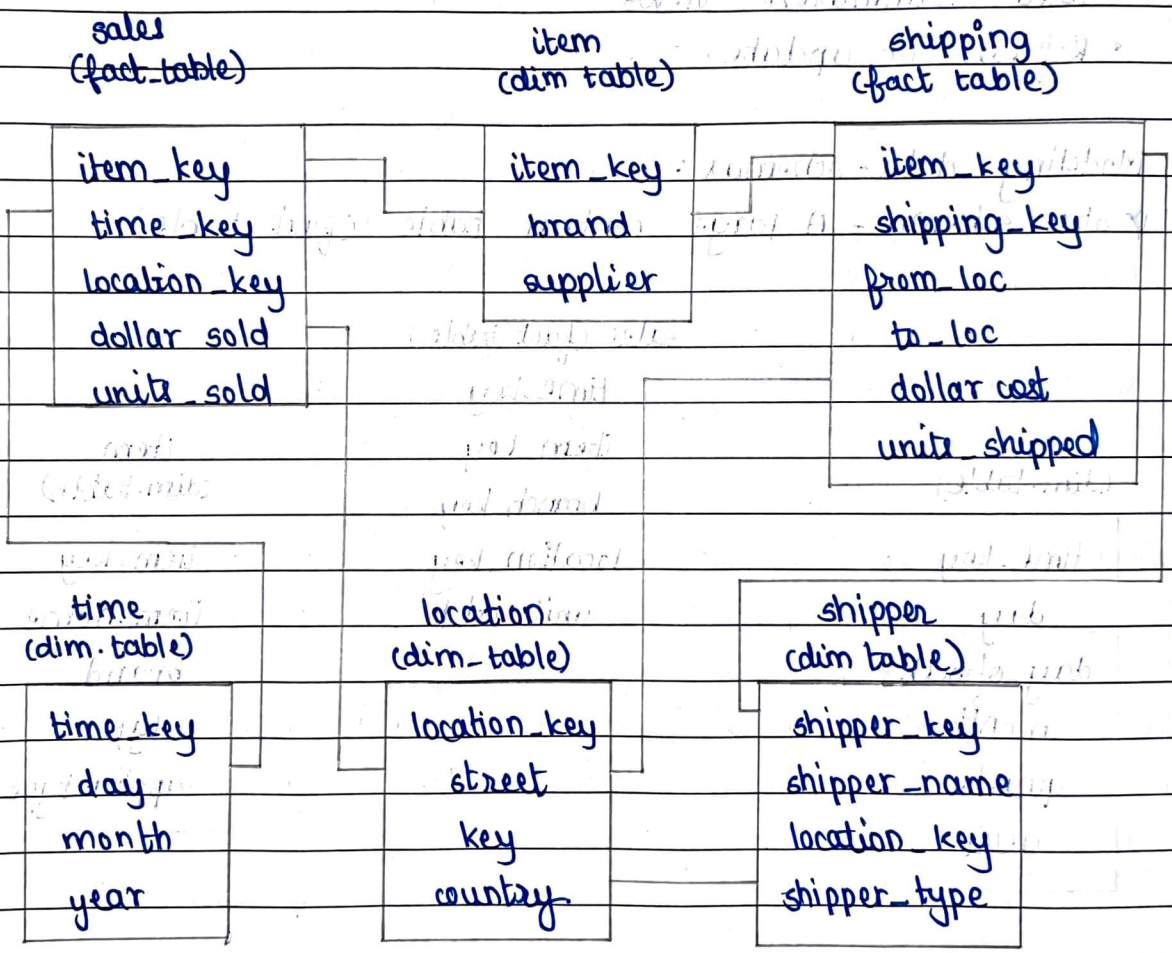
1) star schema - A large central table (fact table)



DATE:

2) Snowflake schema - variant of star schema. Some dim. tables are normalized by further splitting are reduced to redundancy test may increase fetch time.

3) Fact constellation -
collection of star schema i.e. more than a fact table




* Conclusion:
Thus, we successfully installed pentaho pdi-ce-7 ETL tool and performed transformations on sales data.