Assignment No. 1

* Problem Statement:
  For an organization, choose a set of business Processes. design Star/snowflake Schema for analyzing these Processes. Create a fact constellation Schema by combining them. Extract data from diff. data sources, apply suitable transations and load into designation table using ETL tool

* Objective : i.) understanding the basics of Star/ snowflake/ fact constellation Schema.
ii.) Learn how to install & use Pentaho tool.

* Outcome : i.) Pentaho was successfully installed
ii.) Process Schema set was decided & Schema
iii.) ETL was performed on dataset in Pentaho

* Theory :

- ETL Stands for Extract, Transform & load. It extracts data from diff. RDBMS source system, transform data by applying calculation & concatination then load data into data warehouse system. It is loaded in the form of dimention & fact table.

- Extraction

A staging area is req. during ETL load for various reasons.

(1.) The source system are only available for a specific period to extract data. this period of time is less than data load time. this Staging area allows to extract data from source system & keep it in staging area before the time slot.

(2.) Staging area is also req. when you want to get data from multiple source together

- Transformation

In data transformation, you apply a set of functions on extracted data to load it into the target system.

Data which does not req. any transform is known as direct move pass thrw data

Diff transformations on extracted data from the source system can apply, customised to the business decided.

- Load

In this phase, data is loaded into the end target system, either a flat file or a Data warehouse.

⊙ Star Schema ⇒

- Every dimention represented with only one dimention table.
- dimention table should contain a set of attributes
- dimention table joined to fact table using foreign k
- dimention table are not joined to each other
- fact table contains key & measure
- easy to understand, widely supported by BI tools
- optimal disc usage, denormalized structure, faster queries
- dimention tables are not normilized
- cube processing is faster

⊙ Snowflake Schema is an extension of Star Schema, & it adds additional dimentions

- dimention table normilized, which splits data into additional tables
- uses smaller disc space
- reduced query performance due to additional table
- complex database design
- low level data redundancy.

⊙ Pentaho data integration provides the ETL capabilities that facilitates the process of consuming, cleaning & storing data using a uniform & consistent format that is accessable & relevent to end users.

* This tool enables users to -
1) load huge datasets into databases
2) clean data with steps ranging from ver
simple to very powerful complex transformati
3) data warehouse populations, and more

* The PDI client 'Spoon' is a visual, drag
& drop, no-code interface that is used
to perform ETL.

* Conclusion:
1) Pentaho Data integration tool was
successfully installed & used on a Sales
dataset to successfully Extract, transfor
& load in MySQL.
2) A business was picked & a set of
business processes developed & represented
using Snowflack Schema.