Assignment 1

Title : Analyzing and extracting data using ETL tools.

Problem Statements :

For an organization of your choice choose a set of business process. Design star / snow flake schemas for analysing these processes create a fact constellation schema by combining them. Extract data from different data sources, apply suitable transformations and load data into destination tables using ETL tools. For eg. Business organization sales, ordering marketing process

Objective :

1) Implementing of problem statement using ETL tool.

2) Star / snow flake schema for analyzing process

Software and hardware requirements :

2 GB RAM, 500 GB HDD
ETL open source tool pentaho
Mysql
Tomcat with oracle.

Theory :

i) Star Schemas :

The schemas are a way to organize data marts or entire data warehouse using relational databases.

Consider the following sales model represented in star schema

dim_salestype

sales_type_id	int
type value	varchar

dim store

storeid	int
storeadd	varchar
city	varchar
region	varchar

dim_employee

empid	int
name	varchar
birthyear	int

fact_sales

p_id	int
timeid	int
storeid	int
empid	int
price	dec.

dim time

timeid	int
actiondate	int
actionwork	int
actionmonth	int

dim product

pid	int
p_name	varchar
p_type	varchar

A Star Schema

Characteristics of schema

- Every dimension is represented with only one dimensional table
- fact table would contain key & measure
- It is easy to understand and provides optimal disk usage
- Widely supported by BI tools.

2) Snowflake Schema.

- It is an extension of star schema and it adds additional dimensions
- The dimension tables are normalized which splits data into additional table

Characteristics of snowflake schema

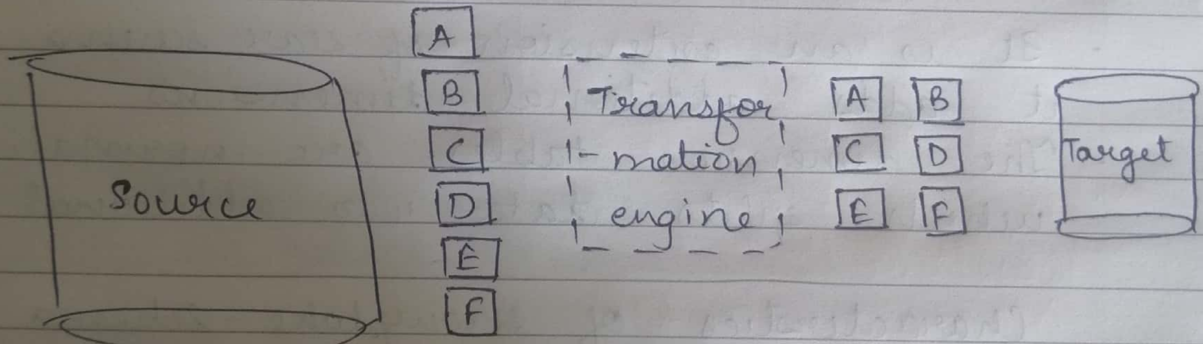
- The main benefit is that it uses small disk spaces.
- Easier to implement a dimension is added to schema.
- Due to multiple tables, query performance is reduced.
- Need to perform more maintainance efforts because of more lookup tables.

3> ETL (Extract, Transform, load)

- ETL is an abbreviation for Extract, Transform and load.

- In this process an ETL tool extracts the data from different RDBMS source systems, then transforms the data by applying calculations concatenations, etc

- In ETL data flows from source to target



List of open sources ETL tools.

- Clover ETL
- Jedox
- Pentaho
- Talend

Test Cases.

Sr No	Description	Expected O/P	Actual O/P
1	Xampp/Apache Server installation	installed successfully	Starts apache & mysql server
2	While installing pentaho make sure to set correct path for environment	success	success
3	Perform transformation on postal codes	success	success
4	Perform transformation on missing zipcodes	Null values removed	success

Conclusion: Thus we have learned to extract data from different data sources, apply suitable transformations and load into destination table using ETL tool.