ASSIGNMENT- 01

Date: 01/08/2020

Title: Analyzing and Extracting data using ETL tools

Problem statement: 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 suitabe transformations and load data into destination table using ETL tools. For example: Business Organization Sales, Ordering, Marketing process.

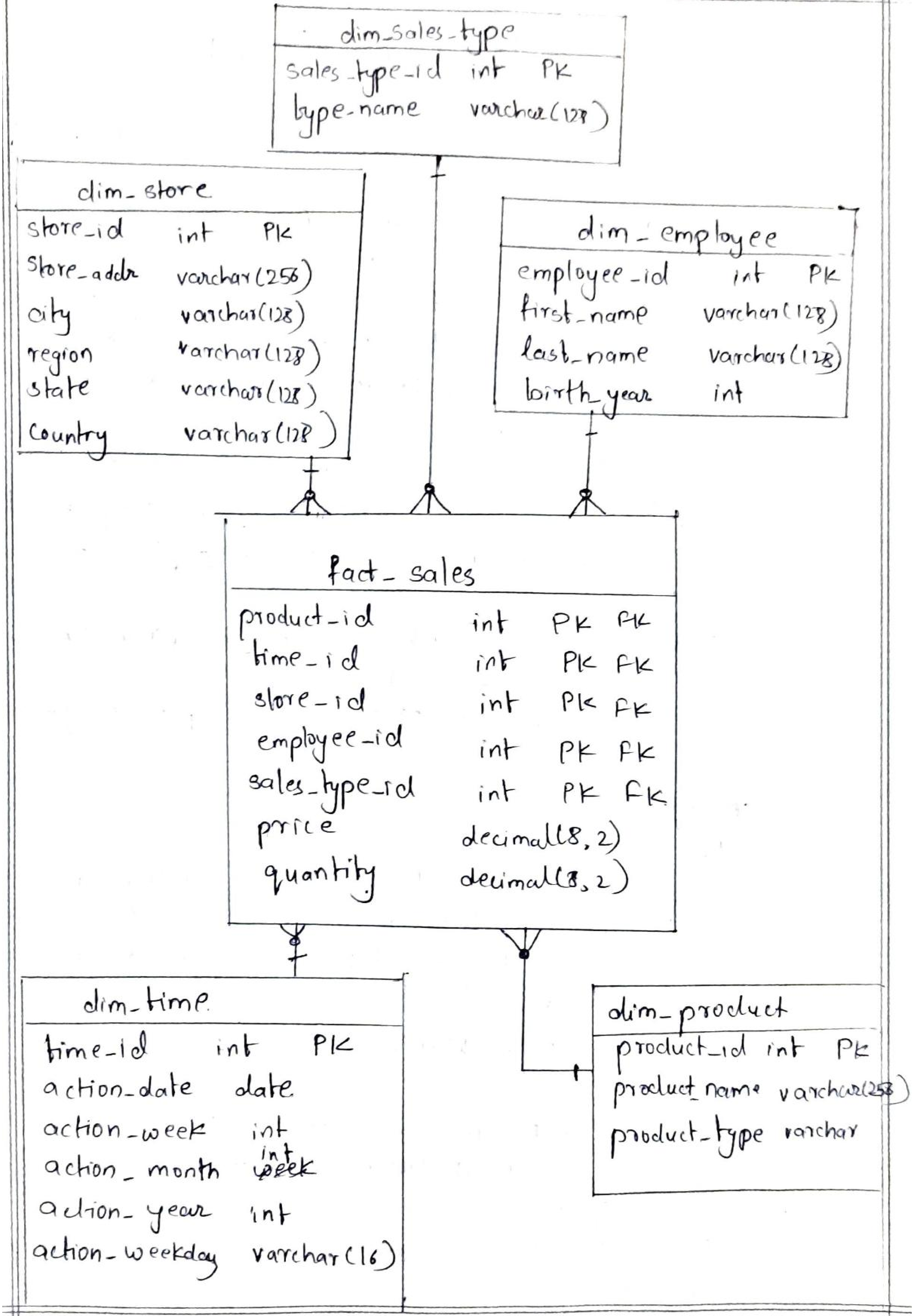
Objective: • Implementation of problem statement using ETL tool.
• star/snow flake schema for analysing process.

S/W & H/W requirements:
- PIV, 2GB RAM, HDD.
- ETL opensource tool Pentaho.
- Tomcat 8.0x with Oracle Java 8.x
- MySQL 5.6 & 5.7 (SQL 92)

Theory: A] 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_sales_type		
sales_type_id	int	PK
type-name	varchar(128)	

dim_store		
store_id	int	PK
Store_addr	varchar(256)	
city	varchar(128)	
region	varchar(128)	
state	varchar(128)	
Country	varchar(128)	

dim_employee		
employee_id	int	PK
first_name	varchar(128)	
last_name	varchar(128)	
birth_year	int	

fact_sales			
product_id	int	PK	FK
time_id	int	PK	FK
store_id	int	PK	FK
employee_id	int	PK	FK
sales_type_id	int	PK	FK
price	decimal(8,2)		
quantity	decimal(8,2)		

dim_time		
time_id	int	PK
action_date	date	
action_week	int	
action_month	int ~~week~~	
action_year	int	
action_weekday	varchar(16)	

dim_product		
product_id	int	PK
product_name	varchar(256)	
product_type	varchar	

A STAR SCHEMA.

→ Characteristics of Star schema:

- Every dimension is represented with the only one-dimensional table.
- Fact table would contain key and measure.
- It is easy to understand and provides optimal disk usage.
- It is widely supported by BI tools.
- The dimension tables are not joined to each other.

B] A 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 a snowflake schema.

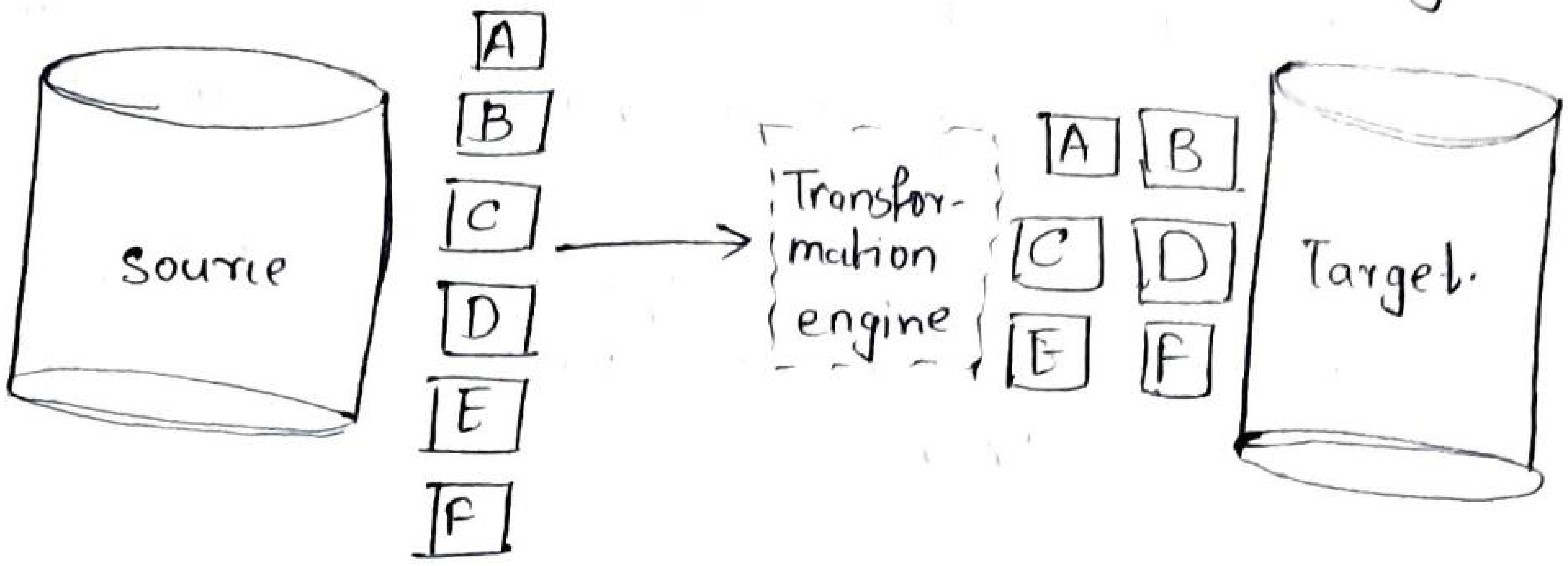
- The main benifit is that it uses a smaller disk space.
- Easier to implement, a dimension is added to schema.
- Due to multiple tables, query performance is reduced.
- Need to perform more maintaince efforts because of more lookup tables

C] 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 and load the Data into data warehouse system.

- In ETL, data flows from source to target.



- ETL is a different method of looking at the tool approach to data movement.
- Instead of transforming data before its written, ETL lets the target system to do transformation.
- The data is first copied to the target and then transformed in place.
- ETL is usually used with no-sql databases like Hadoop Cluster, data appliance or cloud installation

→ List. of open source ETL Tools
- Clover ETL
- Jedox
- Pentaho
- Talend

Test Cases:

Sr. no.	Description.	Expected o/p	Actual o/p
1.	While installion pentaho, make sure to set PENTAHO_JAVA_HOME and		

	PENTAHO_INSTALLERLICENSE_PAT. environment variables	Success	Success
2.	Perform transformation on the postal codes.	Succesfully implemented	Successfully implemented.
3.	Perform transformation on missing_zipcodes.	Null values and unrequired data is removed	Success
4.	Xampp/Apache server installaton	installed successfuly	Starts Apache & MySQL server

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