Assignment - 01 Date: 01/08/2020

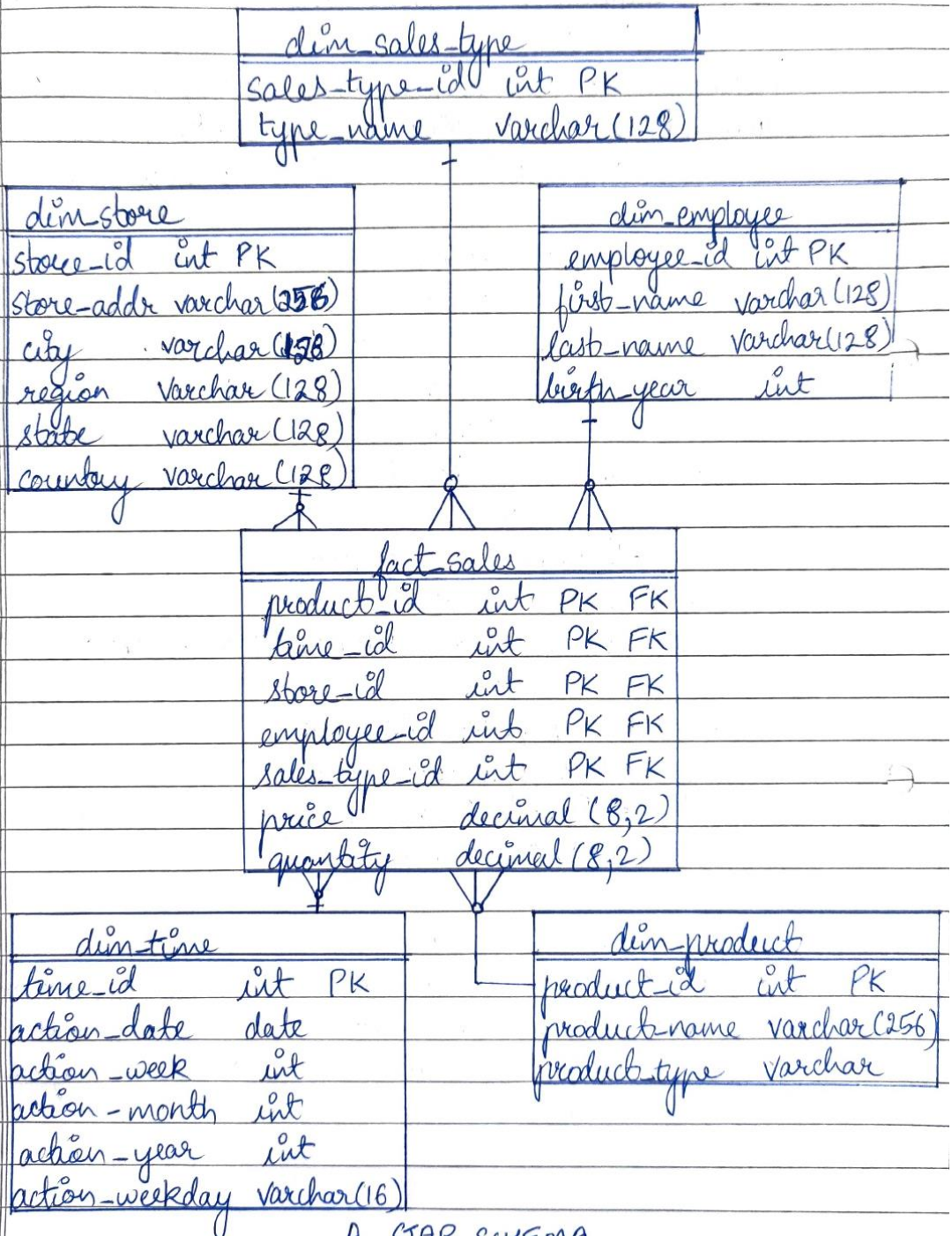
Title: Analyzing and Extracting data using ETL tools

Problem Statement: For an organization of your choice, choose a set of business processes. Design Star/Snow flake schemas for 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 table using ETL tools. For eg. Business Organization sales, ordering, Marketing process.

Objective:
- Implementing of problem statement using ETL tool.
- Star/Snow flake schema for analyzing process

S/W & H/W req:
- PIV, 2 GB 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.



A STAR SCHEMA

→ Characteristics of Star schema:

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

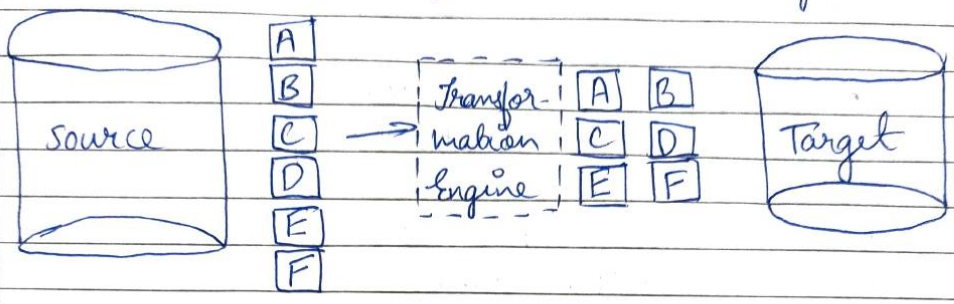
→ Characteristics of snowflake schema.

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

C] ETL (Extract, Transform, Load)

- ETL is an abbreviation for Extract, Transform & 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)	Xampp/Apache Server installation.	installed successfully	starts Apache & Mysql server.
2) & 3)	While installing pentaho, make sure to set PENTAHO_JAVA_HOME & PENTAHO_INSTALLER LICENSE.PAT environment variables	success	success

Sno	Description.	Expected O/P	Actual O/P
2)	Perform transformation on the postal codes	successfully implemented	successfully implemented.
4)	Perform transformation on missing zipcodes	Null values & invalid data is removed	success

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