LP-2.

Assignment 1

Date of Completion:
20.8.2020

Date of Submission:
29.8.2020

Title:- Design star/snow flake schemas for analyzing the process.

Problem Statement:- Design 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 suitable transformations and load into destination tables using an ETL tool. For example: Business Organition: Sales, Order, Marketing Process.

Learning Objective:- To understand star/snow flake schema.
To understand ETL tools.

Learning Outcomes:- Students will be able to use ETL tools to design and also be able to analyze different schemas (snow, star etc)

Software/Hardware Requirement:- An ETL eg Pentaho Data Integration tool.

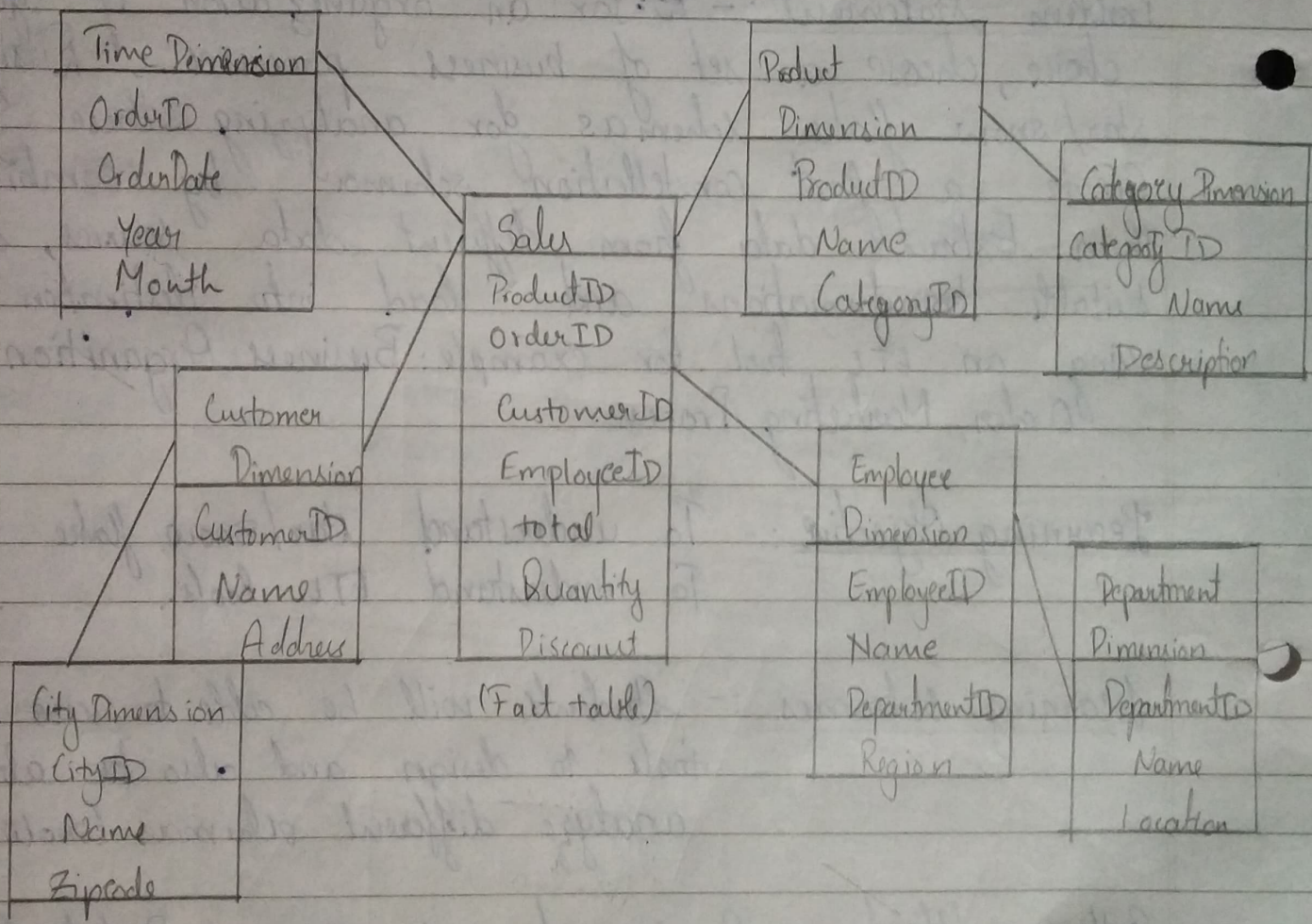
Theory:-

Snowflake schema:-

It is variant of star schema.

The centralized fact table is connected to multiple dimensions.
In the snowflake schema, dimension are present in a normalized form in multiple related tables.
The effect affects only the dimensions tables and not the fact table

Eg.



Dimensions are maintained in normalized form to reduce dependancy. Tables are easy to maintain & save storage space.

Characteristics

1. Use small disk space
2. Easy to implement
3. Performance reduces due to multiple table.
4.

Advantages

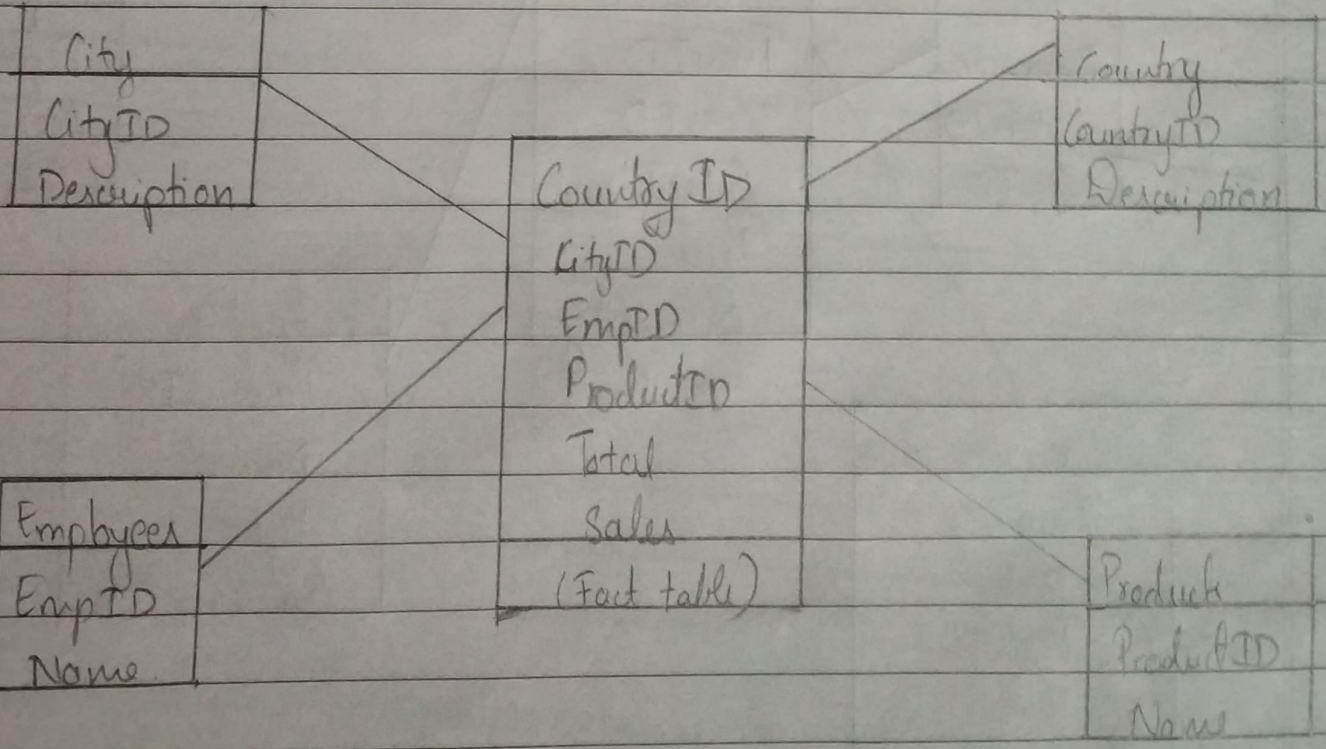
1. Provides structed data which reduces the problem of data integrity.
2. Highly structured data, reduces disk space

Disadvantages:-

Hierarchies should belong to the dimension table only and should never be snowflaked.

Star schema:-

Can have one fact table and many dimension table.
Simplest type of data Warehouse Schema.

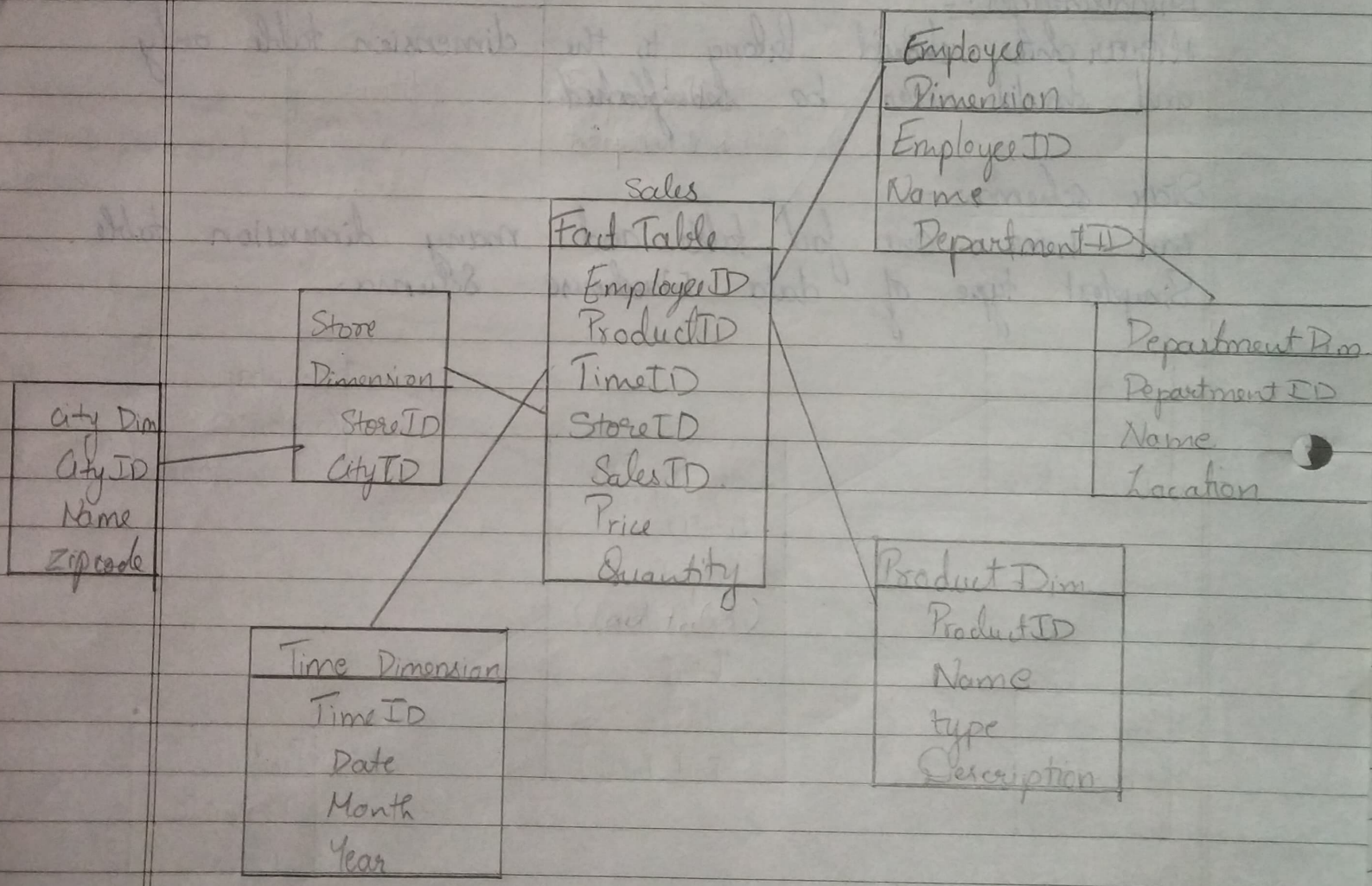


Characteristics:-

1) Only one ~~dimension~~ fact table
2) Dimension table ar not joined to each other
3) Easy to understand
4) Dimension table are not normalized.

Selected Organization

College Snow flake Schema



ETL Tools:-

1) ETL stands for extract tranform Load tool
2) Extract means ~~ext~~ extracting the data from heterogeneous or homogeneous sources into our environment for integeration and generate insights from it.
3) ETL tools extract data from different sources (tables, flat files, etc) and process this data.
4) Transformation phase data is cleansed according to need. Data can be trimmed, appended, filtered, joins can be generated etc
5) In the load phase, final data is loaded into the target DB or into flat files & or in the form of web service.

Eg of ETL tools are Informatica cloud, Ablnitio etc.

Conclusion:- Thus I understood the different schema used (star, snow flake etc), used Pentaho tool for ~~form~~ performing translation.

Terminal File Edit View Search Terminal Help

```
mysql> desc SALES_DATA;
+------------------+------------+------+-----+---------+-------+
| Field            | Type       | Null | Key | Default | Extra |
+------------------+------------+------+-----+---------+-------+
| ORDERNUMBER      | bigint(20) | YES  |     | NULL    |       |
| QUANTITYORDERED  | int(11)    | YES  |     | NULL    |       |
| PRICEEACH        | double     | YES  |     | NULL    |       |
| ORDERLINENUMBER  | int(11)    | YES  |     | NULL    |       |
| SALES            | double     | YES  |     | NULL    |       |
| ORDERDATE        | datetime   | YES  |     | NULL    |       |
| STATUS           | varchar(20)| YES  |     | NULL    |       |
| QTR_ID           | int(11)    | YES  |     | NULL    |       |
| MONTH_ID         | int(11)    | YES  |     | NULL    |       |
| YEAR_ID          | int(11)    | YES  |     | NULL    |       |
| PRODUCTLINE      | varchar(25)| YES  |     | NULL    |       |
| MSRP             | bigint(20) | YES  |     | NULL    |       |
| PRODUCTCODE      | varchar(15)| YES  |     | NULL    |       |
| CUSTOMERNAME     | varchar(50)| YES  |     | NULL    |       |
| PHONE            | varchar(25)| YES  |     | NULL    |       |
| ADDRESSLINE1     | varchar(50)| YES  |     | NULL    |       |
| ADDRESSLINE2     | varchar(50)| YES  |     | NULL    |       |
| CITY             | varchar(25)| YES  |     | NULL    |       |
| STATE            | varchar(25)| YES  |     | NULL    |       |
| POSTALCODE       | varchar(25)| YES  |     | NULL    |       |
| COUNTRY          | varchar(25)| YES  |     | NULL    |       |
| TERRITORY        | varchar(15)| YES  |     | NULL    |       |
| CONTACTLASTNAME  | varchar(20)| YES  |     | NULL    |       |
| CONTACTFIRSTNAME | varchar(20)| YES  |     | NULL    |       |
+------------------+------------+------+-----+---------+-------+
24 rows in set (0.08 sec)

mysql> Select * from SALES_DATA limit 5;
+-------------+-----------------+-----------+-----------------+---------+---------------------+---------+--------+----------+----------+-------------+------+----------
---+---------------------+-----------------+-----------------------------+-------------+-----------+------+-----------+-------------+----------+----------
------+-----------------+
| ORDERNUMBER | QUANTITYORDERED | PRICEEACH | ORDERLINENUMBER | SALES   | ORDERDATE           | STATUS  | QTR_ID | MONTH_ID | YEAR_ID | PRODUCTLINE | MSRP | PRODUCTCO
DE | CUSTOMERNAME         | PHONE           | ADDRESSLINE1                | ADDRESSLINE2 | CITY        | STATE | POSTALCODE | COUNTRY       | TERRITORY | CONTACTLAS
TNAME | CONTACTFIRSTNAME |
+-------------+-----------------+-----------+-----------------+---------+---------------------+---------+--------+----------+----------+-------------+------+----------
---+---------------------+-----------------+-----------------------------+-------------+-----------+------+-----------+-------------+----------+----------
------+-----------------+
|       10107 |              30 |      95.7 |               2 |    2871 | 2003-02-24 00:00:00 | Shipped |      1 |        2 |    2003 | Motorcycles |   95 | S10_1678
   | Land of Toys Inc.    | 2125557818      | 897 Long Airport Avenue     | NULL         | NYC         | NY    | 10022      | United States | NA        | Yu
      | Kwai             |
|       10121 |              34 |     81.35 |               5 |  2765.9 | 2003-05-07 00:00:00 | Shipped |      2 |        5 |    2003 | Motorcycles |   95 | S10_1678
   | Reims Collectables   | 26.47.1555      | 59 rue de l'Abbaye          | NULL         | Reims       | NULL  | 51100      | France        | EMEA      | Henriot
      | Paul             |
```

File Edit View Action Tools Help

Perspective: Data Integ

View | Design

Steps

- Input
- Output
- Transform
- Utility
- Flow
- Scripting
- Lookup
- Joins
- Data Warehouse
- Validation
- Statistics
- Job
- Mapping
- Inline
- Experimental
- Deprecated
- Bulk loading
- History

Transformation1

100%