A

**SEMINAR REPORT**

on

**“VISUALIZATION OF THE THREE SCHEMAS”**

Submitted in partial fulfillment of the requirement

for the assignment of Data Mining Techniques

in Computer Science and Engineering of

Jawaharlal Nehru Technological University, Kakinada

Submitted by

**DMT Batch-1**

Under The Guidance of

**Dr. M. H. M. Krishna Prasad**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**Jawaharlal Nehru Technological University**

**Kakinada District, Andhra Pradesh – 533003**

**2022-23**

**Jawaharlal Nehru Technological University**

**Kakinada District, Andhra Pradesh – 533003**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

CERTIFICATE

**This is to certify that the seminar titled "**VISUALIZATION OF THE

THREE SCHEMAS **" submitted by** DMT Batch-1 **of Semester V is a**

**bonafide account of the work done by him/her under our supervision,**

**during the academic year 2022-2023**

SEMINAR GUIDE HEAD OF THE DEPARTMENT

**ABSTRACT**

Do you think you can derive insights from raw data? It’s possible, of course, but it can be tiresome and not be as accurate as it should be. Wouldn’t the process be much easier if the raw data were more organized and cleaner? Here’s when Data warehousing comes in handy. It is the process of constructing a data warehouse containing essential data. We need to archive and store the data for future use. ETL (Extract, Transform and Load) turns raw data into information.

Through this paper, let’s understand schemas and their types in data warehouse modelling.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| Chapter | Contents | Page No. |
| 1.0 | INTRODUCTION | 01 |
| 2.0 | STAR SCHEMA | 02 |
| 3.0 | SNOWFLAKE SCHEMA | 05 |
| 4.0 | FACT CONSTELLATION SCHEMA | 08 |
| 5.0 | CONCLUSIONS | 10 |
| 6.0 | BIBLIOGRAPHY | 11 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Fig No. | Title | Page No. |
| 2.1 | Structure of star schema | 02 |
| 2.2 | Star schema of a sales data warehouse | 03 |
| 3.1 | Structure of snowflake schema | 05 |
| 3.2 | Snowflake schema of a sales data warehouse | 06 |
| 4.1 | Structure of fact constellation schema | 08 |
| 4.2 | Fact Constellation schema of a sales data warehouse | 09 |
| 5.1 | Comparison between snowflake and star schema | 10 |

CHAPTER I

**INTRODUCTION**

The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for online transaction processing. A data warehouse, however, requires a concise, subject-oriented schema that facilitates online data analysis.

The most popular data model for a data warehouse is a multidimensional model, which can exist in the form of a star schema, a snowflake schema, or a fact constellation schema.

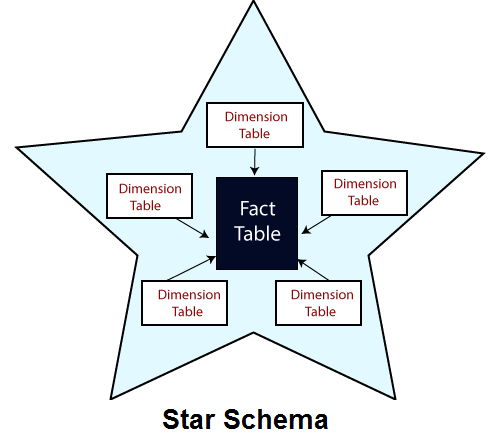
To get a good understanding on the schemas in data warehouse, let us go through few key elements.

1. **Primary Key** – An attribute in a relational database having unique values. There are no duplicate values. We identify each record with its unique value.
2. **Foreign Key** – An attribute in a relational database that links one table to another. It refers to the primary key from another table.
3. **Dimensions** – Dimensions are the column names in a dimension table. Also, dimensions have their attributes sub-divided in the table. We use dimensions as a structured way of describing and labelling the information. Dimension tables are the tables describing dimensions.
4. **Fact Table** – A fact table contains a dimension key from the dimension table and measures. The measures here are to perform calculations for analysis.

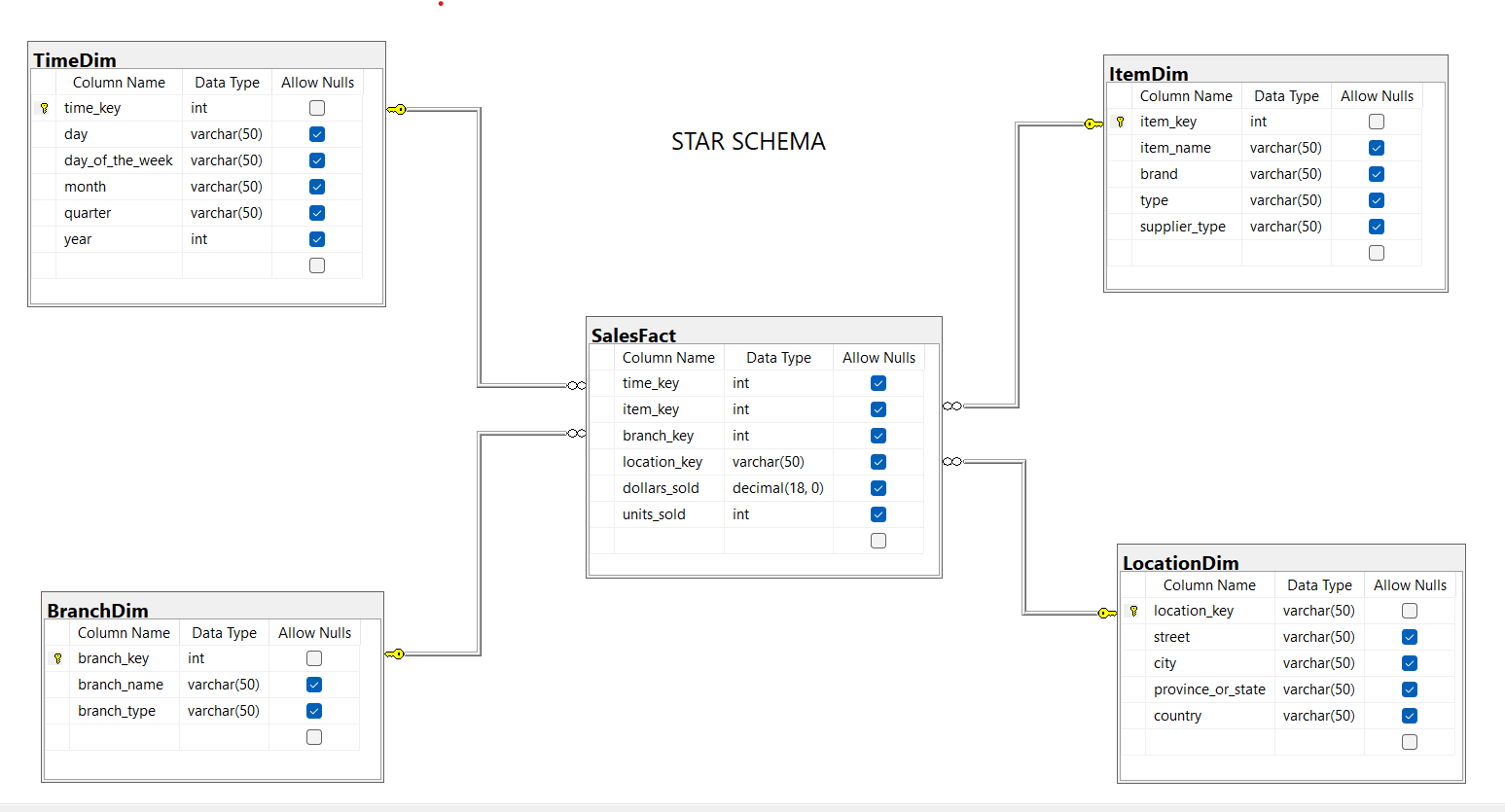
CHAPTER II

**STAR SCHEMA**

* Star schema is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one table indexing any number of dimensional tables.
* The star schema is a necessary cause of the snowflake schema.
* It is also efficient for handling basic queries.It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star’s points.



**Figure 2.1** Structure of star schema.

**Example: Star schema.** A star schema for All Electronics sales is shown in below Figure. Sales are considered along four dimensions: time, item, branch, and location. The schema contains a central fact table for sales that contains keys to each of the four dimensions, along with two measures: dollars sold and units sold. To minimize the size of the fact table, dimension identifiers (e.g., time key and item key) are system-generated identifiers. Notice that in the star schema, each dimension is represented by only one table, and each table contains a set of attributes. For example, the location dimension table contains the attribute set {location key, street, city, province or state, country}. This constraint may introduce some redundancy. For example, “Urbana” and “Chicago” are both cities in the state of Illinois, USA. Entries for such cities in the location dimension table will create redundancy among the attributes province or state and country; that is, (..., Urbana, IL, USA) and (..., Chicago, IL, USA). Moreover, the attributes within a dimension table may form either a hierarchy (total order) or a lattice (partial order).

**Figure 2.2** Star schema of sales data warehouse.

**ADVANTAGES**

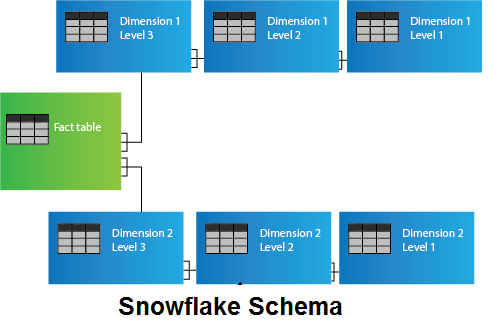
1. **Simpler queries –**  
   join logic of star schema is quite cinch in comparison to other join logic which are needed to fetch data from a transactional schema that is highly normalized.
2. **Simplified business reporting logic –**   
   in comparison to a transactional schema that is highly normalized, the star schema makes simpler common business reporting logic, such as of reporting and period-over-period.
3. **Feeding cubes –**   
   star schema is widely used by all OLAP systems to design OLAPcubes efficiently. In fact, major OLAP systems deliver a ROLAP mode of operation which can use a star schema as a source without designing a cube structure.

**DISADVANTAGES**

1. Data integrity is not enforced well since in a highly de-normalized schema state.
2. Not flexible in terms if analytical needs as a normalized data model.
3. Star schemas don’t reinforce many-to-many relationships within business entities – at least not frequently.

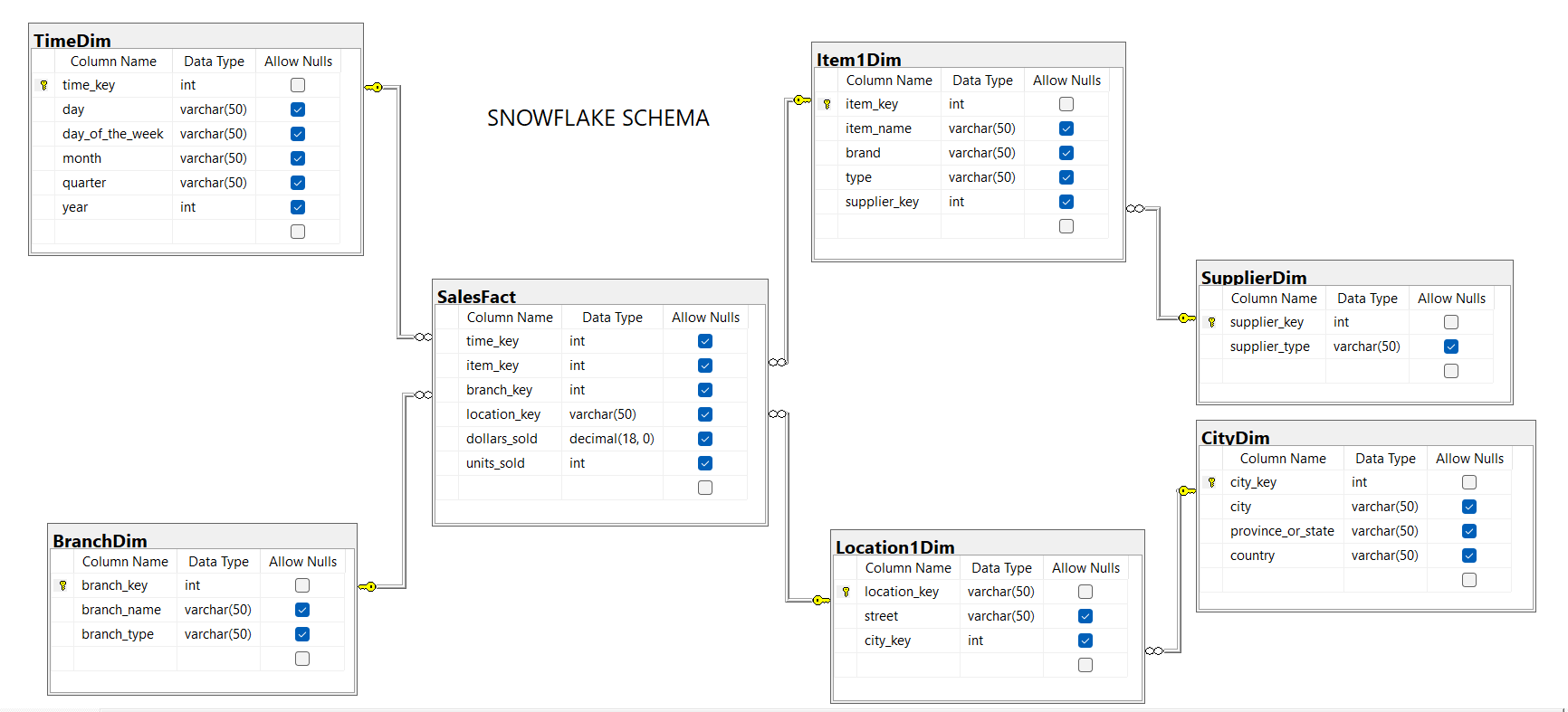
CHAPTER III

**SNOW FLAKE SCHEMA**

* A snowflake schema is equivalent to the star schema. "A schema is known as a snowflake if one or more-dimension tables do not connect directly to the fact table but must join through other dimension tables."
* The snowflake schema is an expansion of the star schema where each point of the star explodes into more points. It is called snowflake schema because the diagram of snowflake schema resembles a snowflake.
* Snowflaking is used to develop the performance of specific queries. The schema is diagramed with each fact surrounded by its associated dimensions, and those dimensions are related to other dimensions, branching out into a snowflake pattern.
* The snowflake schema consists of one fact table which is linked to many dimension tables, which can be linked to other dimension tables through a many-to-one relationship.
* Tables in a snowflake schema are generally normalized to the third normal form.

**Figure 3.1** Structure of snowflake schema

**Example Snowflake schema**. A snowflake schema for AllElectronics sales is given in Figure 3.2. Here, the sales fact table is identical to that of the star schema in Figure 2.2. The main difference between the two schemas is in the definition of dimension tables. The single dimension table for item in the star schema is normalized in the snowflake schema, resulting in new item and supplier tables. For example, the item dimension table now contains the attributes item key, item name, brand, type, and supplier key, where supplier key is linked to the supplier dimension table, containing supplier key and supplier type information. Similarly, the single dimension table for location in the star schema can be normalized into two new tables: location and city. The city key in the new location table links to the city dimension. Notice that, when desirable, further normalization can be performed on province or state and country in the snowflake schema shown in below figure.



**Figure 3.2 snowflake schema of a sales data warehouse**

**ADVANTAGES**

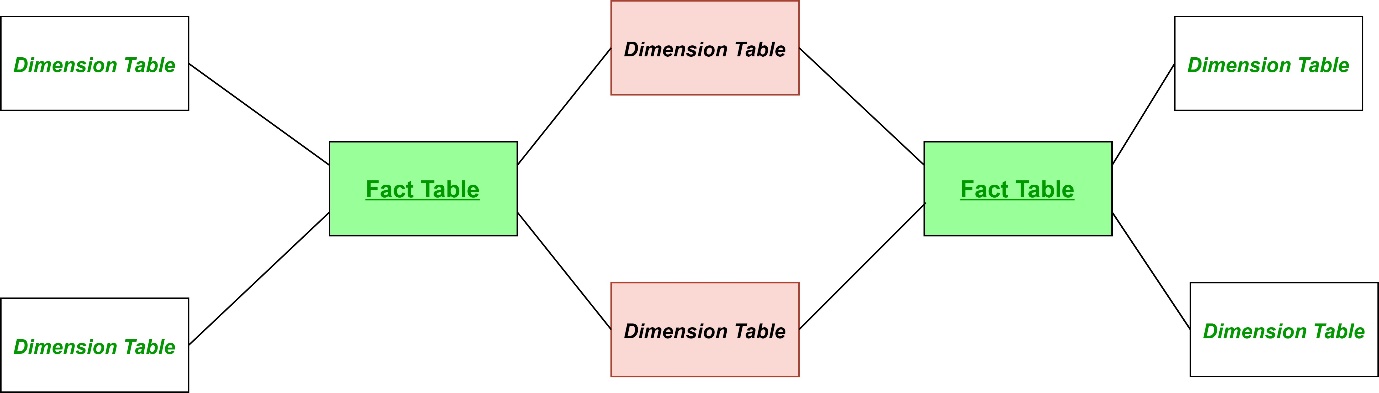
1. The primary advantage of the snowflake schema is the development in query performance due to minimized disk storage requirements and joining smaller lookup tables.
2. It provides greater scalability in the interrelationship between dimension levels and components.
3. No redundancy, so it is easier to maintain.

**DISADVANTAGES**

1. The primary disadvantage of the snowflake schema is the additional maintenance efforts required due to the increasing number of lookup tables. It is also known as a multi fact star schema.
2. There are more complex queries and hence, difficult to understand.
3. More tables more join so more query execution time.

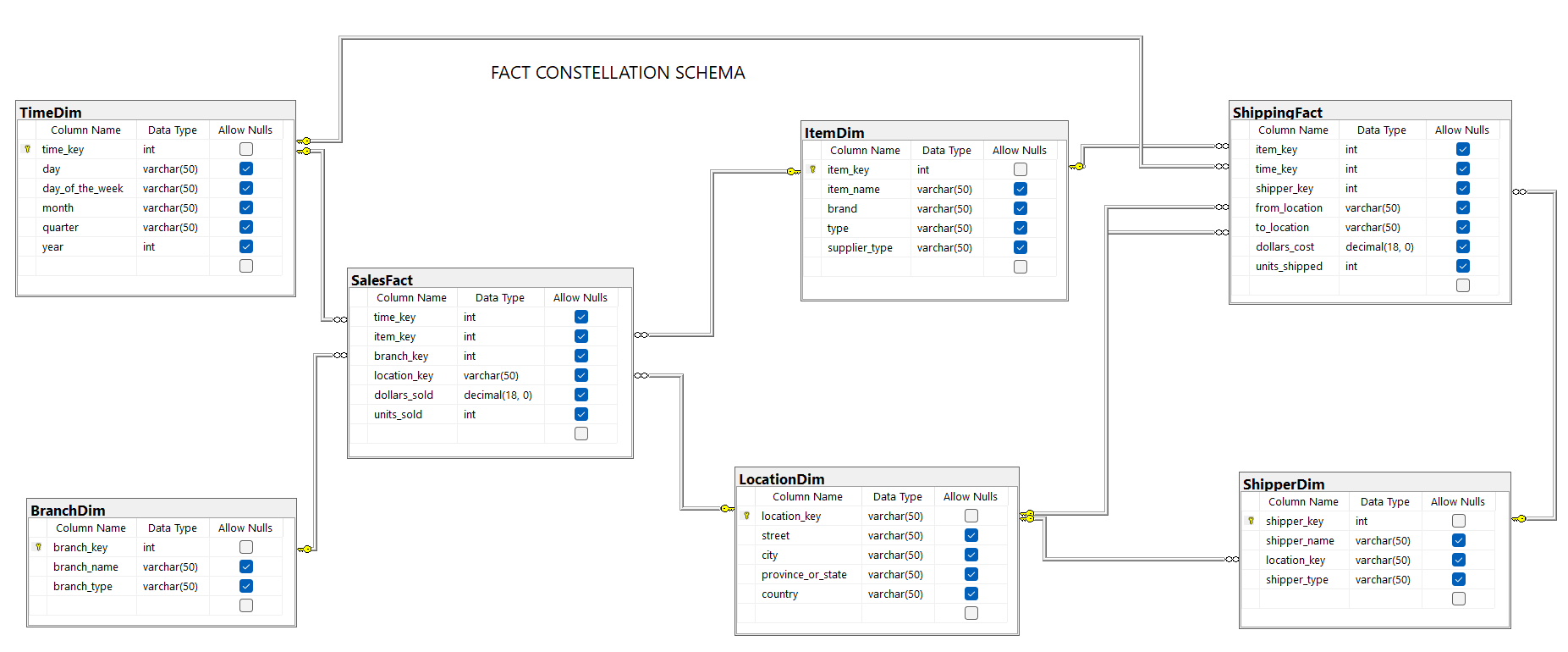
CHAPTER IV

**FACT CONSTELLATIONS SCHEMA**

* Fact constellation is a schema for representing multidimensional model. It is a collection of multiple fact tables having some common dimension tables.
* It can be viewed as a collection of several star schemas and hence, also known as galaxy schema.
*  It is one of the widely used schema for data warehouse designing and it is much more complex than star and snowflake schema. For complex systems, we require fact constellations.

**Figure 4.1 Structure of fact constellations schema**

**Example Fact constellation**. A fact constellation schema is shown in below Figure. This schema specifies two fact tables, sales and shipping. The sales table definition is identical to that of the star schema. The shipping table has five dimensions, or keys—item key, time key, shipper key, from location, and to location—and two measures—dollars cost and units shipped. A fact constellation schema allows dimension tables to be shared between fact tables. For example, the dimensions tables for time, item, and location are shared between the sales and shipping fact tables.



**Figure 4.2 fact constellation schema of a sales data warehouse**

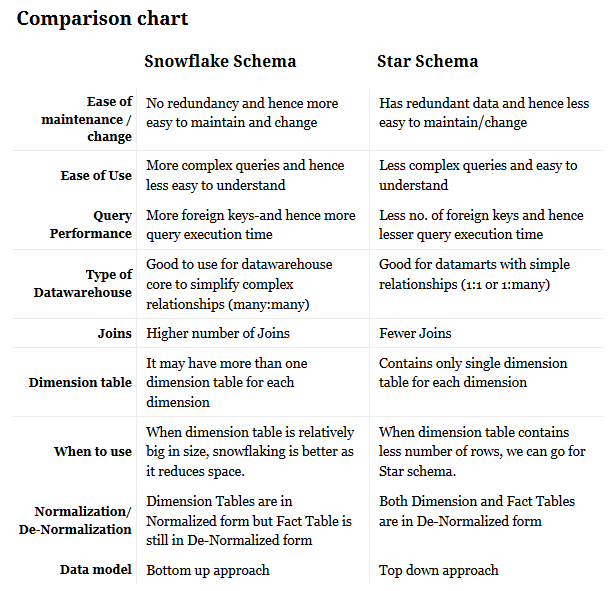
**ADVANTAGES**

* Provides a flexible schema.

**DISADVANTAGES**

* It is much more complex and hence, hard to implement and maintain.

CHAPTER V

**CONCLUSIONS**

**Figure 5.1 Comparison between snowflake and star schema.**

We hope you got a good understanding of different types of Data Warehouse Schemas, along with their benefits and disadvantages from this tutorial.

We also learned how Star Schema and Snowflake Schema can be used, and which schema is to choose between these two along with their differences.

CHAPTER VI

**BIBLIOGRAPHY**

* **Data MINING concepts and techniques** By Jiawei Han, Micheline Kamber, Jain Pei
* <https://www.researchgate.net/>
* <https://www.analyticsvidhya.com/blog/2022/06/what-are-schemas-in-data-warehouse-modeling/>