I J R B A T, Issue (XI) Vol (II) May 2023 : 222-230

A Double-Blind Peer Reviewed & Refereed Journal

OPEN ACCESS

e-ISSN 2347 – 517X

Original Article

Reg. No. : MH-NGP/0001G1

INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY

© VMS RESEARCH FOUNDATION www.ijrbat.in

COMPARISON OF DATA MODELS IN DATA WAREHOUSE

Pooja D. Kavishwar[1] and S. R. Pande[2]

[1,2]Department of Computer Science, Shivaji Science College, Congress Nagar, Nagpur, Maharashtra, India

Communicated : 18.01.2023 | Revision : 22.02.2023 & 07.03.2023 Accepted : 29.03.2023 | Published : 30.05.2023

ABSTRACT:

Data warehousing (DW) provides an approach in converting operational data into beneficial and reliable information to assist the decision making process in any organization. In this paper, a process to implement a DW is stated. This paper is used to elaborate and compare various conceptual, logical and physical design models for data warehousing. This comparison is done to find which of the conceptual and logical data models are more appropriate for implementing a data warehouse which helps in enhancement of the business and system.

Keywords :- Data warehouse, Data warehouse Models, Comparison of models.

INTRODUCTION :

Information is an asset to any organization. Earlier, traditional database systems, called operational or transactional databases were used to store data. But analysis of data for decision making was not possible as they are mostly normalized, they perform poorly for complex queries that need to join many relational tables or to aggregate large volumes of data.

To overcome this use of the Data warehouse has started. Data warehousing provides an approach in reconstructing operational data into useful and reliable information to support the decision making process. Data analysis techniques like data mining and multidimensional analysis are possible. According to W.H. Inmon, “A Data Warehousing (DW) is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management’s decision making process” [1], [2]. Data warehousing process contains pulling out of data from varied data sources, cleaning, filtering, transforming and storing data into a common structure that can be easily accessed and used for reporting and analysis purposes. The data warehouse can be created or updated at any time, queried for any information with minimum disruption to operational systems.

To design a data warehouse data modelling is required which is a process of creating a data model for databases used in designing data warehouses. There are two data modelling techniques that are pertinent in a data warehousing environment: ER modelling and Multidimensional modelling.

ER Modelling :

ER modelling produces a data model of the particular area of interest, using two basic concepts: entities and the relationships between those entities. The ER model is an abstraction tool because it can be used to understand and simplify the confusing data relationships in the business world and complex systems environments.

***Entity* :** An entity is described to be a person, place, thing, or event of interest to the business. An entity also represents a class of objects, which are things in the real world that can be observed and classified by their properties and characteristics.

***Relationship* :** Relationship illustrates association among the entities in a model. It is designated grammatically by a verb, such as

http://doi.org/10.29369/ijrbat.2023.02.1.0032

owns, belongs, and has in a data model. The relationship between two entities can be defined in terms of the cardinality which is the maximum number of instances of one entity that are related to a single instance in another table and vice versa. The viable cardinalities are: one-to-one (1:1), one-to-many (1:M), and many-to-many (M:M).

Multidimensional Modelling :

Multidimensional Modelling applies three basic concepts: measures, facts, and dimensions which signifies the requirements of the business user in the context of database tables.

Fact and measure - The fact models the subjects, the events or the phenomena that the decision makers of the organization need to analyse. Every fact is categorized by one or several measures representing the analysed indicators. A measure is a numerical property of a fact which describes a quantitative attribute relevant to the analysis [4][5].

Dimension is the attributes corresponding to the information that makes the measures of activity vary, i.e., it is the axis of the analysis. The dimension is organized in hierarchies to enable analysis of the measures at various levels of detail [6].

Both ER and Multidimensional modelling can be used to generate an abstract model. Using these two techniques Data modelling is possible.

TYPES OF MODEL

There are primarily three different types of data models:

***Conceptual*:** This Data Model describes WHAT the system includes. Business stakeholders and Data Architects creates this model. The purpose is to organize and define business concepts and rules.

***Logical*:** Describes HOW the system should be executed regardless of the DBMS. Data Architects and Business Analysts creates this model. The objective is to develop technical map of rules and data structures.

***Physical*:** This Data Model defines HOW the system will be executed using a particular DBMS system. DBA and developers creates this model. The purpose is actual application of the database.

Conceptual design models

Conceptual design modelling develops a formal, complete, abstract design based on the user requirements. User requirements are translated into an abstract The goal of creating a conceptual schema is to translate user requirements into an abstract representation understandable to the user that is independent of implementation issues, but is formal and complete, so that it can be transformed into the next logical schema without ambiguities [7].

StarER Model This model combines star structure with constructs of entity relationship (ER) model. Star structure has one fact table and around that fact table arranged are set of smaller dimension tables. The fact table is linked to all the dimension tables by one to many relationships [2], [8], [9]. The prime constructs of ER model are - the entity sets portraying real-world objects, the relationship sets portraying associations among objects and, the attributes representing properties of entity or relationship set [15]. StarER model has the following constructs [15]: fact set, entity set, relationship set, attribute.

Multidimensional ER (ME/R) Model is an enlargement of the ER Model. There are few additional elements [11]:

A special entity set also known as dimension level

Two special relationship sets connecting dimension levels are

a special n-ary relationship set known as the 'fact' relationship set

a special binary relationship set known as the 'rolls-up to' relationship set

Object Oriented Multidimensional (OOMD) Model OOMD modelling approach is based on

http://doi.org/10.29369/ijrbat.2023.02.1.0032

UML. In OOMD model, dimensions and facts are symbolized by dimension classes and fact classes. OOMD approach uses a generalization-specialization relationship to categorize entities that contain subtypes [12].

Dimensional Fact (DF) Model is a set of tree structured fact schemas. Elements of DF models are facts, attributes, dimensions and hierarchies [13].

Comparison Of Conceptual Design Models

Property 1 (Additivity of measures): The numeric value in a fact table that is more flexible is an additive measure. For each dimension summation is possible. DF, starER and OOMD support this property. ME/R model captures only static data structure. No operational aspect can be applied with ME/R model.

Property 2 (Many-to-many relationships with dimensions): Many dimensions linked with many other dimensions. StarER and OOMD support this property. Many-to-many relationships are not supported by DF and ME/R models.

Property 3 (Derived measures): These are calculated measures or logical measures. Only OOMD model includes derived measures as part of their conceptual schema other than any other model.

Property 4 (Nonstrict and complete classification hierarchies): Although DF and ME/R can define certain attributes for classification hierarchies; starER model can define exact cardinality for nonstrict and complete classification hierarchies. OOMD can represent nonstrict and complete classification hierarchies.

Property 5 (Categorization of dimensions - specialization/ generalization): Classification of dimensions indicate type of data stored. All conceptual design models except DF support this property as they mainly support one to one hierarchy.

Property 6 (Graphic notation and listing user requirements):
Graphical information is provided by all modelling techniques which helps designers in conceptual modelling phase. ME/R model also provides both state diagrams and a basic set of OLAP operations to model system’s behaviour and user requirements, respectively. OOMD provide complete set of UML diagrams and defines OLAP functions to specify user requirements.

Logical design models

Data Warehouse logical design contains the definition of structures enabled for an efficient access to information. A conceptual schema representing the information requirements, the source databases, and non-functional requirements are taken into account by The designer builds relational or multidimensional structures. Flat schema, Terraced schema, Star schema, Starflake Schema, Star Cluster schema, Snowflake schema, Fact Constellation schema etc are Logical Models. Among them, most commercially used models are star schema, snowflake schema and fact constellation schema. Therefore, only these three models are primarily considered in this study.

Fact Constellation Schema A fact constellation schema consists of a set of star schemas with hierarchically linked fact tables. The links between the various fact tables provide the ability to “drill down” between levels of detail [14], [8].

Star Schema It is the basic structure for a dimensional model. A set of smaller dimension tables and a fact table is arranged around the fact table. All the dimension tables and fact table are linked together by one to many relationships. It contains measurements which may be aggregated in various ways [2], [8], [9].

Snowflake Schema A snowflake schema is a modified version of star schema. All its

http://doi.org/10.29369/ijrbat.2023.02.1.0032

hierarchies explicitly shown, and dimension tables do not contain denormalized data [2], [8].

Comparison Of Logical Design Models

Efficiency is the most important factor in DW modelling because many queries access large amounts of data that may involve multiple join operations [8]. A star schema is usually the most efficient design for two reasons. First, a design with denormalized tables as it needs fewer joins. Second, most optimizers understand star schemas and can generate cost-effective "star join" operations. A fact constellation schema is a set of star schemas with hierarchy. A fact constellation schema may require more join operations on fact tables. Also, a snowflake schema will require more joins on dimension tables. But in few cases where the denormalized dimension tables in star schema becomes huge, a snowflake schema may be the most efficient and used design approach. In terms of usability, few more advantages may be considered for star schema design approach. The star schema is the cleanest structure among the three schemas. Because a star schema has least number of tables, users need to execute fewer join operations which makes it easier to formulate analytic queries. It is easier to understand star schema compared to other two schemas. While considering reusability, the snowflake schema is more reusable than star and fact constellation schemas.

Dimension tables in a snowflake schema do not contain denormalized data. This makes it simpler to share dimension tables between snowflake schemas in a DW. In star schema and fact constellation schema design approaches, dimension tables are denormalized and this makes it inconvenient to share dimension tables between schemas. In terms of flexibility, a star schema is more flexible in accommodating changes in user requirements, as all dimensions are equivalent in terms of providing access to the fact table. Table 2 summarizes the comparison of the three logical design models in terms of quality factors. Although snowflake schema is not very efficient compared to star schema and fact constellation schema because data is fully normalized and extracting data from many tables requires more time; use of snowflake schema for the implementation of the DW is more beneficial. One advantage of storing data in normalized table is that redundancy is minimized and therefore data inconsistency problem will not arise. Another reason for choosing the snowflake schema is that, the sample OLTP database prepared as the data source for DW is completely formed of normalized tables and therefore using snowflake schema in the design of the DW is more applicable and easier to implement

Physical Model:

The physical design model is a procedural step for converting the data into an actual database. Thus, the physical data model is designed from the midlevel data model and by expanding the midlevel data model keys and physical characteristics of the model can be included. Thus, the physical data model looks like a series of tables called relational tables.[3] The physical database tables are ready to be casted into the concrete physical database design with one last design step of optimization of performance characteristics. With the data warehouse, the first and foremost step in design is deciding on the granularity and partitioning of the data. After this, a variety of other physical design activities are embedded into the design. At the heart of the physical design the usage of physical I/O (input/output) are considered. Physical I/O is the activity that brings data in the computer from storage or sends data to storage from the computer. Thus, this states an overview of physical design which is used while designing a Data Warehouse.

http://doi.org/10.29369/ijrbat.2023.02.1.0032

CONCLUSION :

This paper has stated two techniques for designing a Data warehouse models. Which has then extended in defining three structures in which Data warehouse models are accommodated.

All these structures containing model design schemas are individually described with examples and then compared to each other. With the findings in this paper, widely accepted conceptual design models for DW design are inferred. OO design model is significantly better than the other design approaches. Inference of the comparison done between conceptual designs result states that OOMD is the best suited model as it supports all properties used for comparison. OOMD also supports rich set of diagrams which is defined and base for conceptual design to model all the business requirement. Snowflake schema may be the most effective design approach where deformalized dimension tables used in start schema is huge. This comparative study is usual in wisely using model designs to best suit the requirements of the business or user. Thus increasing the efficiency and reusability of Data warehouse.

REFERENCES:

Connolly T., and Begg C., Database Systems, 3rd Edition, Addison-Wesley, ISBN: 0-201-70857-4, 2002.

Han J., and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann, ISBN: 1-55860-489-8, 2000.

Inmon W.H., Building the Data Warehouse, 3rd Edition, John Wiley, ISBN: 0-471-08130-2, 2002.

Ravat F, Teste O, Zurfluh G (2001) Mode´lisation multidimensionnelle des syste`mes de´cisionnels. In EGC, pp 201–212

Kimball R (1996) The data warehouse toolkit practical techniques for building dimensional data warehouses. Wiley, New York

Abello A, Samios J, Saltor F (2001) Understandding analysis dimensions in a multidimensional object-oriented model. In: Proceedings of the international workshop on design and management of data warehouses (DMDW), Interlaken, Switzerland, pp 4-1–4-9

Phipps C., Davis K., “Automating Data Warehouse Conceptual Schema Design and Evaluation”, DMDW’02, Toronto, Canada, 23-32, 2002.

Moody D. L. and Kortink M. A. R., “From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design”, Proceedings of the Second Intl. Workshop on Design and Management of Data Warehouses, DMDW 2000, Stockholm, Sweden, June 5-6, 2000.

Ferguson N., “Data Warehousing”, International Review of Law Computers & Technology, Volume 11, Number 2, pages 243-249, 1997.

Tryfona N., Busborg F., Christiansen J. G., “starER: A Conceptual Model for Data Warehouse Design”, Proceeding of the ACM 2nd International Workshop Data Warehousing and OLAP (DOLAP99), Kansas City, USA, Page.3-8, 1999.

Sapia C., Blaschka M., Höfling G., Dinter B., “Extending the E/R Model for the Multidimensional Paradigm”, Proceeding 1st International Workshop on Data Warehousing and Data Mining (DWDM 98), Springer-Verlag, Vol. 1552, Page.105- 116, 1998.

Trujillo J., Palomar M., “An Object Oriented Approach to Multidimensional Database Conceptual Modeling (OOMD)”, Proceeding 1st International Workshop

http://doi.org/10.29369/ijrbat.2023.02.1.0032

on Data Warehousing and OLAP (DOLAP98), USA, Page.16-21, 1998.

Golfarelli M., Maio D., Rizzi S., “The Dimensional Fact Model: A Conceptual Model For Data Warehouses”, International Journal of Cooperative Information Systems (IJCIS), Vol. 7, Page.215-247, 1998.

Lujan-Mora S., Trujillo J., Song I., “Multidimensional Modeling with UML Package Diagrams”, 21st International Conference on Conceptual Modeling (ER2002), Tampere, Finland, Page.199-213, 2002. [8] Martyn T., “Reconsideri

Tryfona N., Busborg F., Christiansen J. G., “starER: A Con- ceptual Model for Data Warehouse Design”, Proceeding of the ACM 2nd International Workshop Data Warehousing and OLAP (DOLAP99), Kansas City, USA, Page.3-8, 1999.

Table 1 . Comparison of Conceptual design models

Property	DF	starER	ME/R	OOMD
1	✓	✓	X	✓
2	X	✓	X	✓
3	X	X	X	✓
4	X	✓	X	✓
5	X	✓	✓	✓
6	✓	✓	✓	✓

Table 2. Comparison of logical design models

	Star Schema	Fact Constellation Schema	Snowflake Schema
Efficiency	High	High	Moderate
Usability	High	Moderate	Moderate
Reusability	Low	Low	High
Flexibility	High	High	Moderate
Redundancy	High	High	Low
Complexity	Low	Moderate	Moderate

http://doi.org/10.29369/ijrbat.2023.02.1.0032

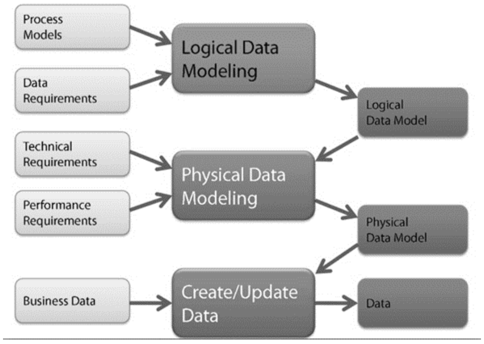

Figure 1 . Types of Models

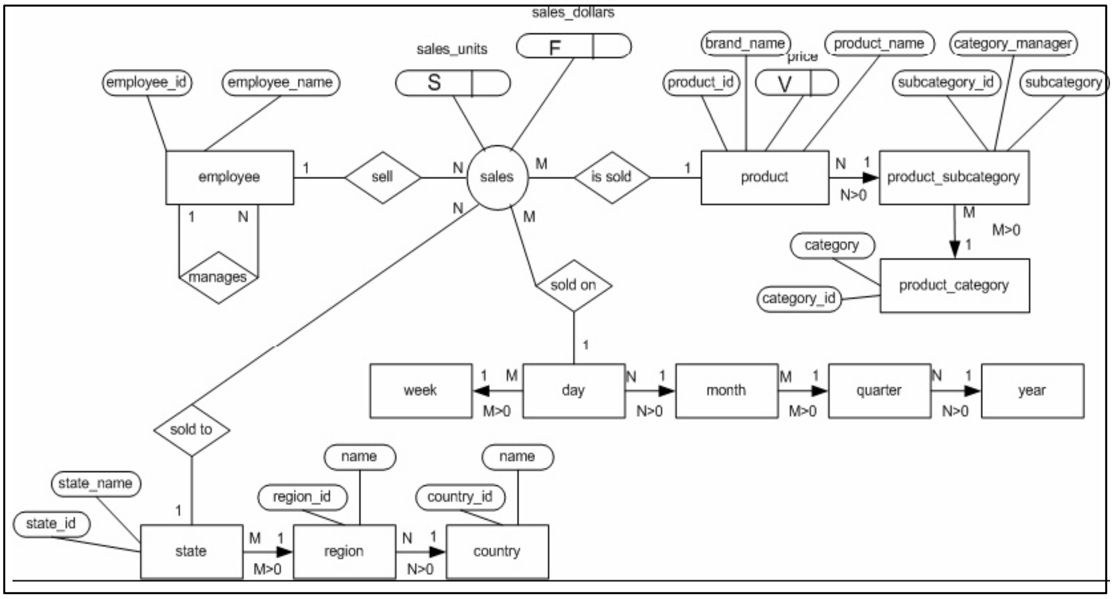

Figure 2 . starER Model

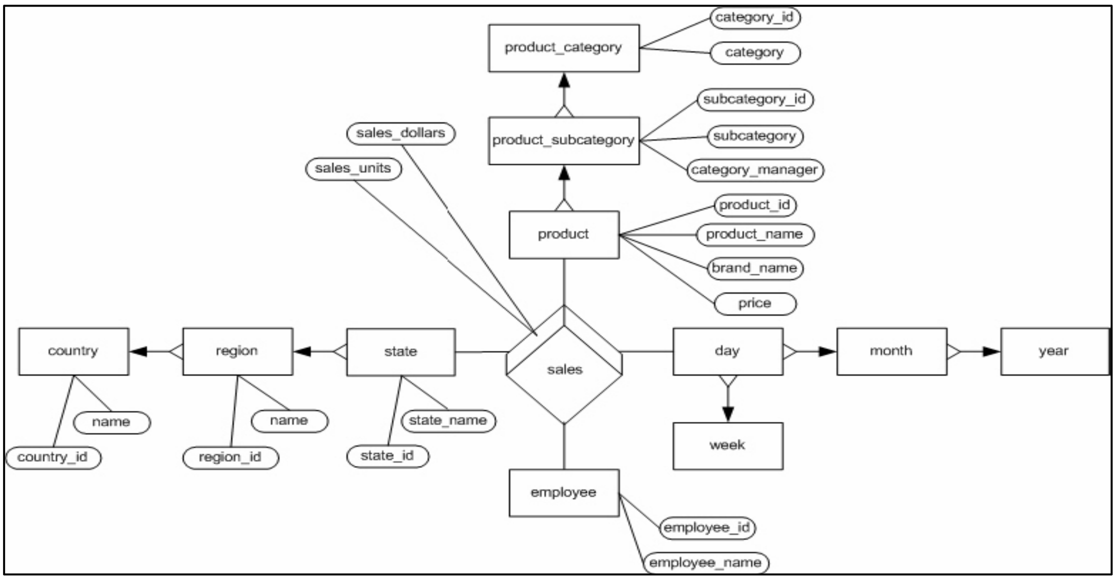

Figure 3 . ME/R Model

http://doi.org/10.29369/ijrbat.2023.02.1.0032

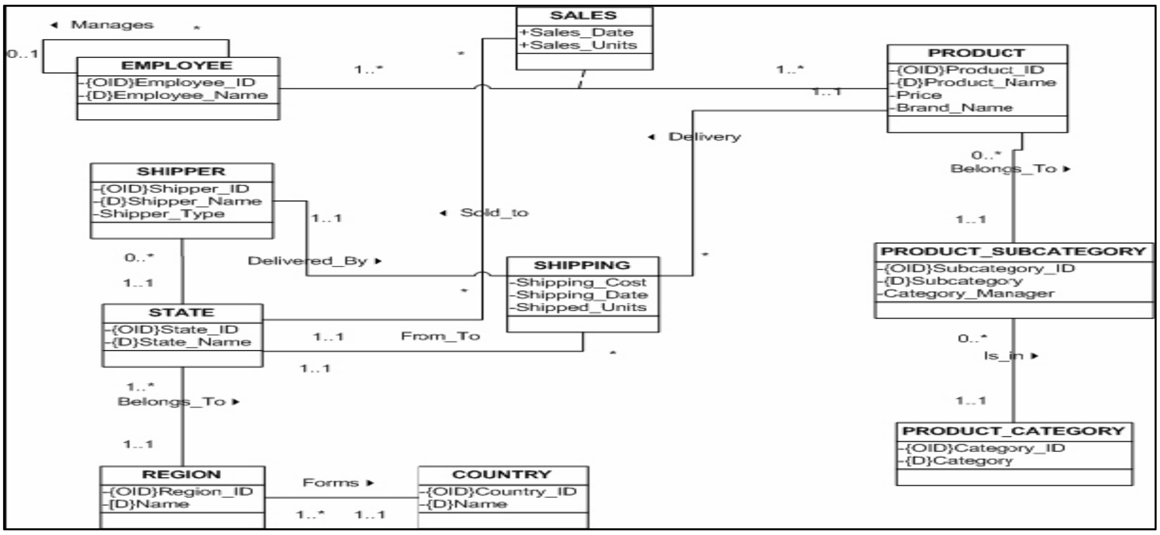

Figure 4 . OOMD Model

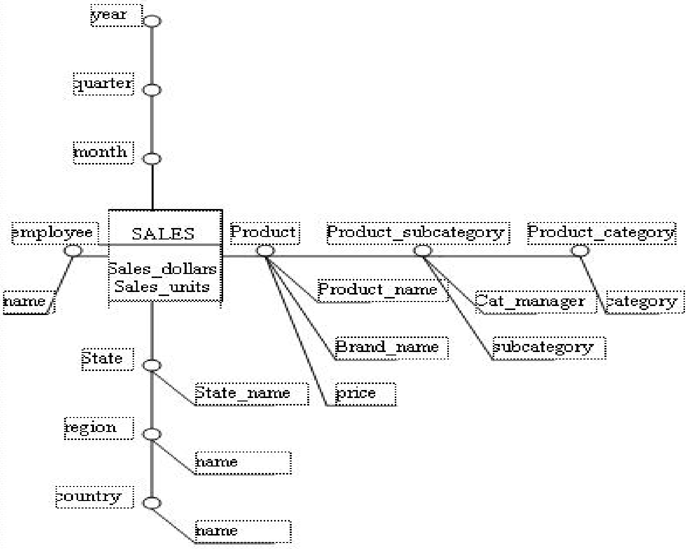

Figure 5 . DF Model

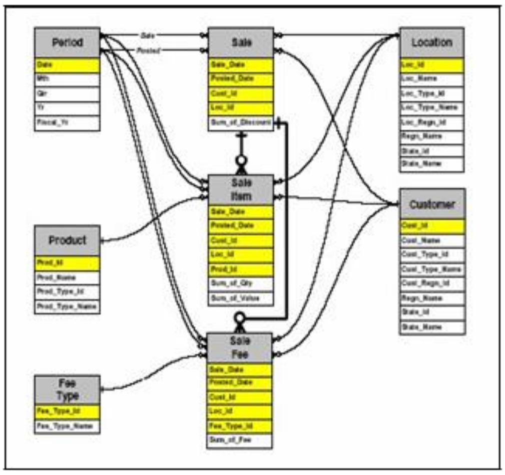

Figure 6 . Fact Constellation Schema

http://doi.org/10.29369/ijrbat.2023.02.1.0032

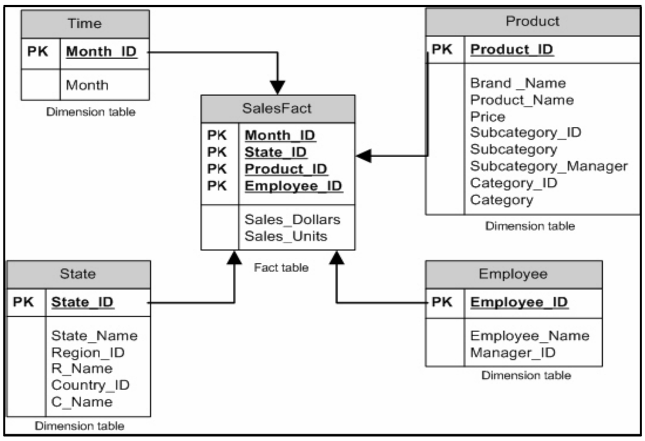

Figure 7 . Star Schema

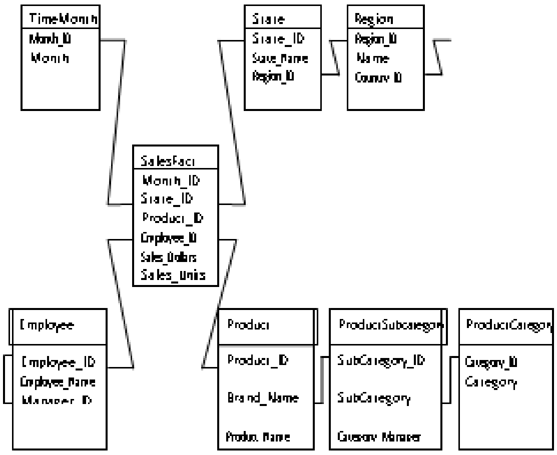

Figure 8 . Snowflake Schema

http://doi.org/10.29369/ijrbat.2023.02.1.0032