**SCHEMA DESIGN**

**MULTIFACT SNOWFLAKE SCHEMA**

**INTRODUCTION:**

Advancements in technology have set the onset of better understanding of biology and the behaviour of diseases. Large datasets that contain knowledge about genetics help in understanding and discovering new facts about diseases. Genetic data tends to be complex and poses challenges in mining, storage and management. While data warehousing and OLAP have been successfully applied to the business domain, it is clear that direct transfer of these technologies to biology is fraught with difficulties[1]. This leads to the need for designing logical models that are adaptable to biomedical data.

**PROBLEM DESCRIPTION:**

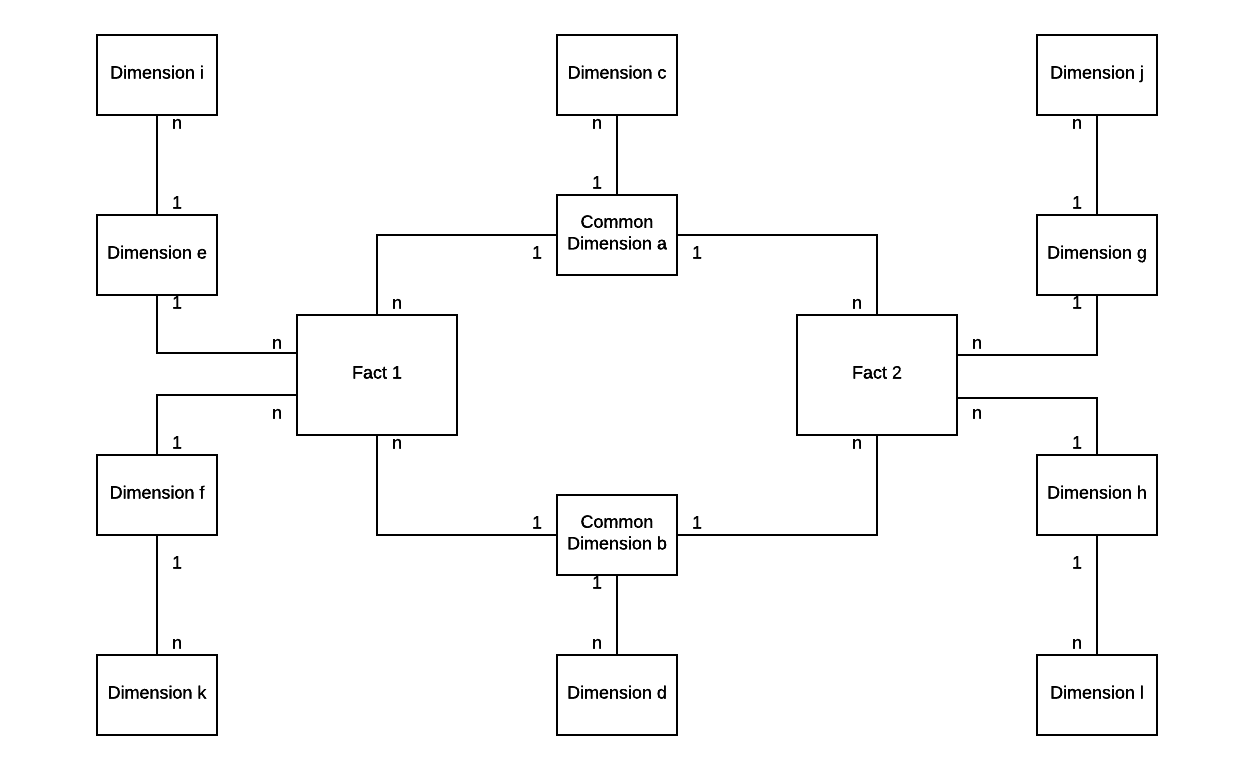
With the increasing use of biomedical data for research and historical data analysis there is a need for flexible representation of this data. Although there are several generic schemas that can be used for this purpose, they do not address the issues that arise when dealing with medical data sets. A database schema that specifically addresses the needs of the biomedical field requires creation. Clinical and genomic data tend to have changing number of dimensions due to the continuous evolution of this data. They also have many-many relationships among facts and dimensions and these relationships tend to be uncertain. This data also has changing validity with new advances and hence requires time series management. It also tends to be incomplete or imprecise. These problems need to be addressed when creating schemas for biomedical data.

**OUR APPROACH:**

The logical data model that we designed borrows modelling strategies from Snowflake Schema and Multi-fact Star Schema. It explores the multi-fact asset of biomedical data. It also tries to overcome most of the challenges posed by clinical data.

***Basic model:***

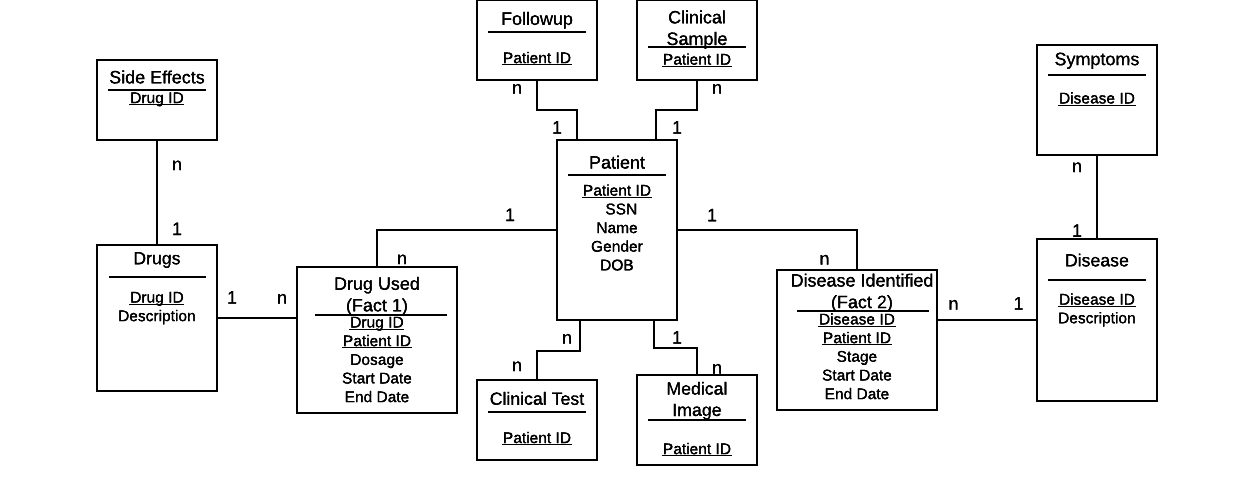
Our basic model has multiple facts. There are various dimensions associated with these fact tables. At least a single dimension is common between the fact tables and they connect the fact tables. This schema can be extended with the requirement of new dimensions to be added by connecting the dimension with the respective fact table and only altering that fact table. Since each dimension can have various other dimensions associated with it, a snowflake schema structure is obtained.



*Figure 1: Multifact Snowflake Schema*

***Case study:***

We use this schema to design logical model for the clinical data space.



*Figure 2: Multifact Snowflake Schema for Clinical Data Space*

***Challenges addressed:***

With the design of this schema we try to address the challenges that biomedical data poses. Many-to-many relationships are handled with the fact tables. Since there are multiple Fact tables each representing one domain of biomedical data the problem of sparse data does not arise. Since our model is easily extensible it works well with changing dimensions of the biomedical data. Since fact tables are specific to the dimensions it is easy to add or remove entries and hence works for incorrect or imprecise data.

**CONCLUSION:**

The logical model that we designed takes features of the snowflake schema and uses the multi fact attribute to enhance it. It also addresses the issues that arise when dealing with biomedical data in a fairly simple manner. The fact tables between the dimensions make it much more flexible in comparison with other models.

**REFERENCES:**

[1]: BioStar models of clinical and genomic data for biomedical data warehouse design

Liangjiang Wang and Aidong Zhang

http://www.cse.buffalo.edu/faculty/azhang/cse601/IJBRA.pdf