SCHEMA DESIGN FOR DATA WAREHOUSE

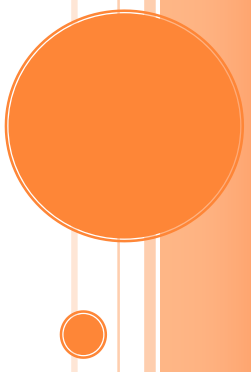
CSE 601 Data Mining

Homework 1

Ankit Jain

Deepak Veerupapuram

Milky Sahu



Schema Design for data warehouse

Introduction:

With the rapid advances in the computerization of data, medical science could rest as one of its most needy beneficiary. Today computer technologies have begun to mature and have opened to us, a vast range of applications. Storing a large amount of information in a central location (databases) could open the door to maintaining complete medical information, personalized medical histories as well as other technologies that could take medical science to new heights. A large data warehouse could lead us to newer approaches of diagnosis not as common as with the keeping of paper files.

Some of the **advantages** of a medical data warehouse could be as follows:

1. Health care providers and insurance companies could use information networks to share electronic medical records. These data banks would help with the reduction of paperwork, in billing, identifying the most cost-effective treatment, and to fighting against false claims.

2. A person's medical information would be immediately available to doctor. Therefore in case where a new doctor sees a patient, he/she would have the patient's entire medical history at their fingertips. Included in this information could be lifesaving information that would be invaluable to the attending doctor.

3. The creation of such a data warehouse would also allow researchers to follow certain diseases as well as to patients' responses to their treatments. This information could be valuable to drug companies for research purposes only.

4. The creation of these databases would allow for better organization and more legibility of medical files.

5. Since elaborate security systems can be developed to monitor these medical databases, electronic records may actually be more secure than paper records.

Major characteristics of clinical and genomic data:

1. Complex data structure with many potential dimensions.
2. Often many-to-many relationships between facts and dimensions.
3. Uncertain relationships between fact and dimension objects.
4. Some clinical measures require advanced temporal support for time validity.
5. Incomplete and/or imprecise data very common.

The above characteristics of biomedical data should be considered as the critical requirements of the multidimensional data model for the biomedical data warehouse. The existing multidimensional models do not fully support these requirements.

We have designed a logical schema that would address all these requirements.

A conventional star or snowflake schema has problems to model the clinical data space. We have designed a schema in which we have lowered the grain of Medical records fact table. This is done by employing two fact tables namely Medical records fact table and Billing fact table. Using this structure, allows us to use two fact tables in different ways.

When we perform disease-related analysis, the Medical Records fact table can be used. When we perform billing–related analysis, the Billing fact table can be used. When we analyze billing related to disease, both fact tables will be used. A billing event is joined to disease through the billing key.

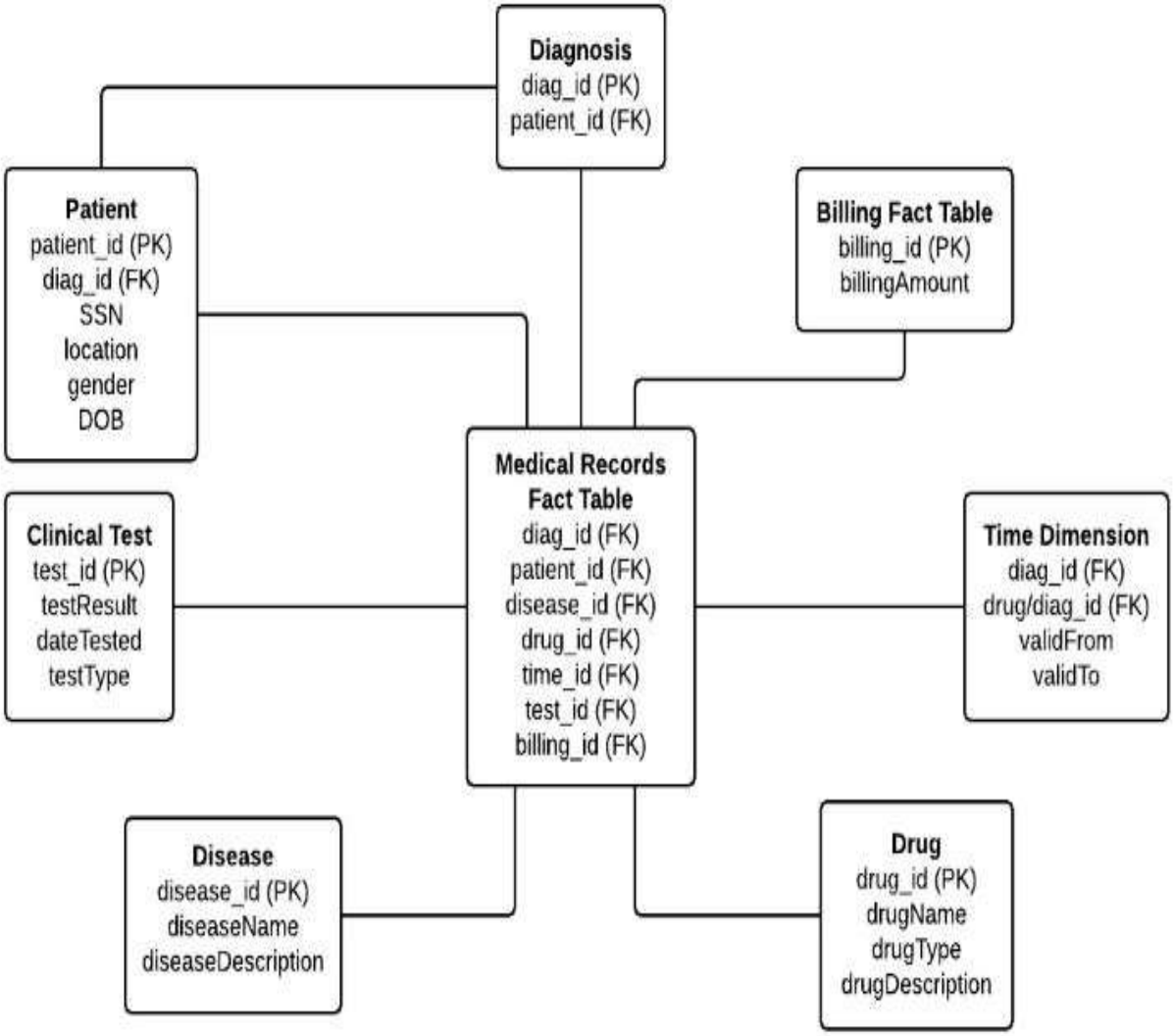
Advantages of this approach:

1. Many to many relationship is resolved by separating data into two fact tables.
2. View is on a single table.
3. Less joins and thus fast queries.
4. Support temporal storage for time validity.
5. Support extensibility of schema when a new dimension is introduced.

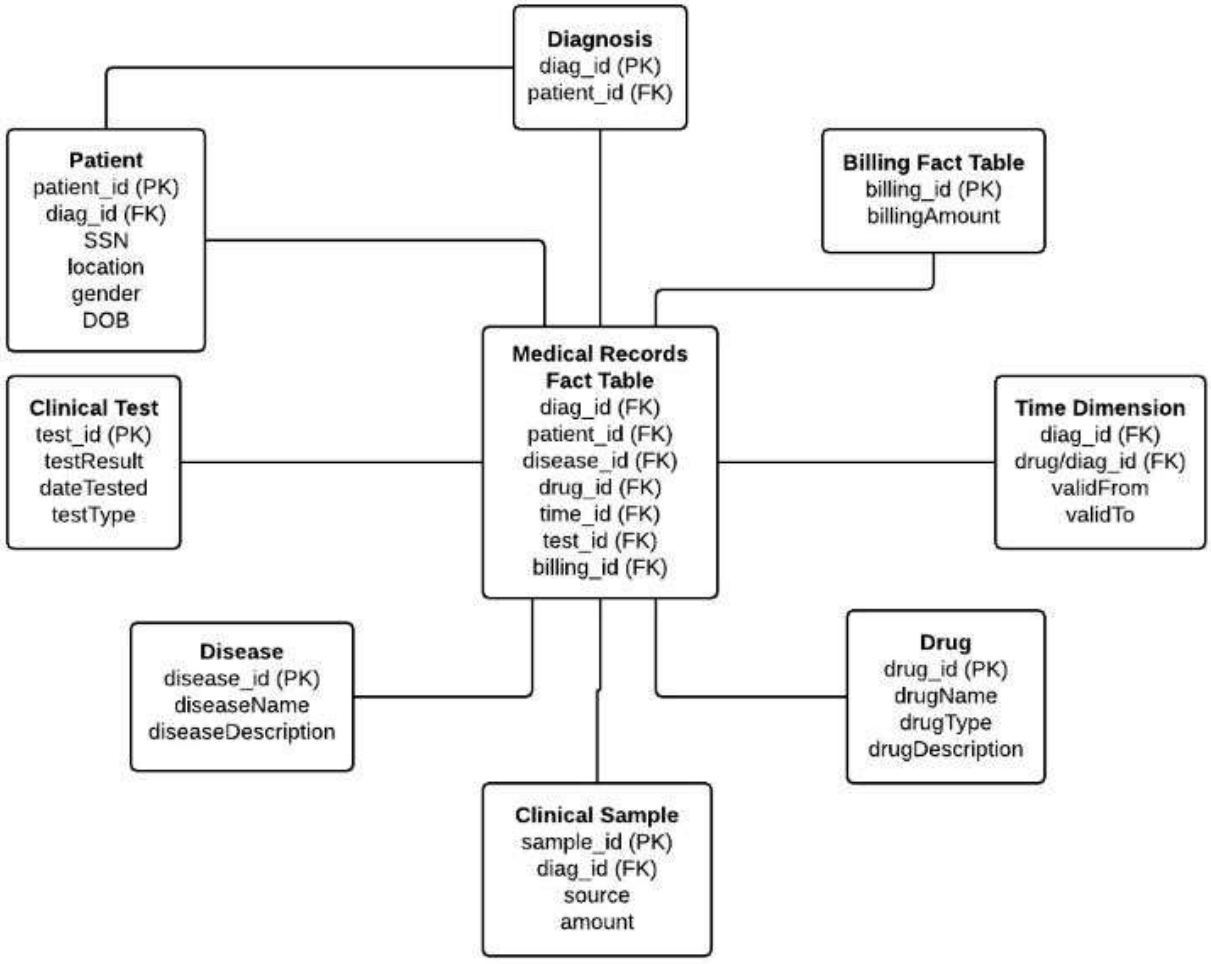
Disadvantages of this approach:

In this approach, the patient medical record would contain redundant data about the patient due to the granularity of the table, which is at the diagnosis level while billing is recorded only once per event in its own fact table. The size of the fact table will increase depending upon how many diagnoses are stored in the fact table.

Schema Diagram:



If a new dimension (Clinical Sample) gets introduced, then



References:

http://www.academia.edu/977976/An_analysis_of_many-to many_ relationships _between _fact _and dimension tables in dimensional modeling

http://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/

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