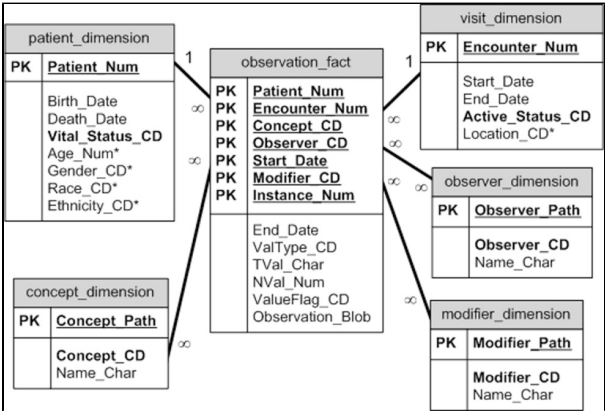# 711. Background

A core feature of the i2b2 platform is a generic database schema based on the Entity-Attribute-Value concept (EAV) [Nadkarni 1997]. It facilitates the storage of heterogeneous and time-varying biomedical data in a unified and stable data model. In it, a central fact table ("OBSERVATION_FACT") is joined with several dimension tables (e.g. "PATIENT_DIMENSION", "CONCEPT_DIMENSION", ...) in a classic data warehouse star schema:

One advantage of this approach is that new data elements can simply be added by including definitions in the relevant dimension table(s) (e.g. CONCEPT_DIMENSION) and adding the respective data rows to the fact table. No changes to the database schema itself are required. This generic approach and the resulting simplicity of importing data into i2b2 have contributed to its high uptake in the biomedical community.

However, an important drawback of the EAV concept is that all fact data elements are stored in a single, large table, which is not fully conducive to classic index-based optimizations available in relational databases. While in a classic, normalized database queries for specific data elements can be focused on individual relevant tables, in an EAV schema there is only one table containing everything, leading to performance impacts with large datasets.

Biomedical data is highly heterogeneous, especially regarding the volume of various data elements within the overall dataset. E.g. in a clinical data warehouse, data volume is often dominated by laboratory findings (dozens per encounter) whereas diagnosis codes are much scarcer (often less than 10 per encounter). Searching for a diagnosis in an EAV fact table entails scanning through a large volume of irrelevant data items (e.g. lab findings). While table scans can be optimized by setting up appropriate indices, their benefits are limited in an EAV context. However, relational database platforms provide additional optimization features (e.g. partitioning) which can be exploited.