# DEVELOPMENT OF PATIENT CARE MODEL CORRESPONDING TO INDIAN HOSPITALS


Enrol. No. -122502

Name of Student -Shivani Singh

Name of supervisor(s) -Dr.Dipankar Sengupta



May-2014

Submitted in partial fulfilment of the Degree of

Master of Technology

In

Computational Biology


**DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS**

**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,**

**WAKNAGHAT**

# TABLE OF CONTENTS

# CERTIFICATE

This is to certify that the work titled "**DEVELOPMENT OF PATIENT CARE MODEL CORRESPONDING TO INDIAN HOSPITALS**" submitted by "**SHIVANI SINGH**" in partial fulfilment for the award of degree of Masters of Technology in Compuational Biology, to Jaypee University of Information Technology, Waknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor ……..

Name of Supervisor ……..

Designation ……..

Date ……..

## ACKNOWLEDGEMENT

***"To speak gratitude is courteous and pleasant, to enact gratitude is generous and noble, but to live gratitude is to touch Heaven"***

As I conclude my project, I have many people to thank; for all the help, guidance and support they lent me, throughout the course of my endeavour. First and foremost, I am highly indebted to my supervisor, **Dr**.**Dipankar Sengupta**, who has guided me through thick and thin. I deem it a privilege to be a student doing research under Dr. Dipankar Sengupta who has endeared himself to his students and scholars.

Secondly, I pay my most sincere thanks to **Prof. (Dr.) R.S. Chauhan**, Head of Department, Department of Biotechnology and Bioinformatics, for providing me with an opportunity and facilities to carry out the project. I also thank **Ms. Somlata Sharma** (Bioinformatics Laboratory In-charge) for her assistance and valuable contribution.

I am indebted to all those who provided reviews and suggestions for improving the results and topics covered in my project, and extend my apologies to any one whom i have failed to recognize in my efforts.

_____

Signature of Student
Name
Date

# SUMMARY

In hospitals large amount of clinical data is being generated everyday. This data cannot be analyzed and evaluated easily. It consists of many hidden patterns or relationships which help in discovering knowledge about the clinical data,which can become hurdles to the patients in one way or the other.

So,for this Patient care model can be developed which helps in improving the safety and quality of care for the patients.It involves creation of dimensional model and a data warehouse which will help in storing all of the patient's data in one place so that it can be managed easily.

As, the clinical data sets are very large, so the time consumed to analyze these data sets will also be very large which will take much time of the patients. In order to reduce this time factor, critical factors which are the cause of increase in time factor are properly identified,by developing a proper data warehouse for storing the data of the patient care model developed with the help of dimensional modelling. Data warehouse is used for reporting and analysis purpose and creates a central repository of the data.

The data is uploaded from the various operational sources which in my case are the files stored in the csv format and then passed to the Kettle for ETL process i.e Extract,transform and load process,where this ETL will use staging,data integration and the access layers to perform various functions.

Design of data warehouse is leads to the formation of dimension model. Physical model is created using Erwin tool which will show the type of data data warehouse will be storing.It consists of dimension and fact tables where columns of dimension tables are linked or related to the fact table with the help of foreign keys.

| | |
|---|---|
| _____ | _____ |
| Signature of Student | Signature of Supervisor |
| Name | Name |
| Date | Date |

## LIST OF FIGURES-

# CHAPTER 1

# INTRODUCTION

With the advancement of time more and more technologies have been used for the care of the patients in the hospitals, use of computational techniques have led to the more ease both for the patients as well as for the doctors and experts. Nowadays use of many electronic machines have reduced the work of doctors as these machines help in identifying the disease and what appropriate measures should be taken in order to cure the disease. This will help in saving the time of both patients as well as of experts.

This is the era of machines as they reduce the effort as well as time of the human in performing any action. Human is totally dependent on machines and life seems impossible without the use of machines.

Hospitals in abroad are more advanced as compared to Indian hospitals, as patient care still needs more effort corresponding to the Indian hospitals. Clinical datasets are very time consuming because of many critical factors involved in it. On identification of rate limiting or these critical factors with the help of data mining techniques the time consumed can be reduced which will be help in decision making process. Decision making process is an important process of implementation as it helps in selecting the required course of action from several alternative processes. There is a term called Business intelligence (BI) which refers to the technology associated with the integration and analysis of collected information. In a health care system or hospital setting, having access to large amounts of information -- whether clinical or financial -- can lead to better evidence-based decision making. Traditionally, information in health systems has been compiled in static, text-based memoranda. BI helps shape that information into visual data that provides a basis for evidence-based decision making in many different departments.

In recent years, health care's adoption of BI systems is on the rise. This is due to continued implementation of electronic health record (EHR) systems, the storing of clinical data in more discrete formats and a variety of federal mandates. The U.S. Department of Health & Human Services requires organizations to submit specific clinical quality reports as part of the meaningful use incentives program, as well as additional quality measures through the Physician Quality Reporting System (PQRS) and The Joint Commission.

The motivation for data mining and analysis does not stop there, though. As health care shifts toward the accountable care organization (ACO) model and other pay-for-performance

initiatives, many organizations will be required to provide proof of improved patient outcomes. This would be directly tied to their reimbursements.

In a typical hospital, there are several areas where BI tools can be used to drive evidence-based decision making. Productivity, efficiency, financial performance and customer service are covered below. The use of business intelligence for clinical analysis is covered in a separate tip, as is an examination of common health care data analysis methods.

So, for supporting this decision making process many decision support systems are developed which help in planning and management of the process. While dealing with the clinical datasets clinical decision support system(CDSS) will be taken into account.

### 1.1 CDSS

CDSS stands for clinical decision support system which can be defined as the use of the computer to bring forward the relevant knowledge to support on the health care and well being of a patient. Clinical decisions mean those that support on the management of health and health care of an individual person (the patient). Support means encouraging of rather than the making of decisions. Pertinent knowledge means the selection of knowledge that is directly relevant to the specific patient. Timing of the support may be different for different CDSS; ease of accessibility may also vary for the clinicians to access. Most of the CDSS are stand alone systems or a part of computer based patient record system. CDSS also vary on the type of the information provided.

### 1.2 FEATURES OF CDSS

The main aim of CDSS is to make the clinical data about a patient easily accessible and to encourage the development of optimal problem solving and decision making process. Selection of the pertinent knowledge or processing of the data to create pertinent knowledge is the main or the primary task of the computer. Selection can be made based on the patient specific data.

Another feature is that this selection and the processing of the data can be done with the help of some inferencing processes. Finally the result of CDSS is to make some recommendations.

### 1.3 WHY CDSS?

It has number of important benefits including-

- Increased quality of care and enhanced health outcomes.
- Avoidance of errors and adverse events.
- Improved efficiency, cost-benefit, and provide patient satisfaction.

It is a sophisticated health IT component and requires computable biomedical knowledge, person-specific data, and a reasoning or inferencing mechanism that combines knowledge and data to generate and present helpful information to clinicians as care is being delivered.

### 1.4 TYPES OF CDSS

*Knowledge based and Non-knowledge based-*

*Knowledge based-*

Knowledge based systems are those which uses a knowledge base in order to solve complex problems. Many of today's knowledge-based CDSS arose out of earlier expert systems research, where the aim was to build a computer program that could simulate human thinking. Medicine was considered a good domain in which these concepts could be applied. The main aim of these CDSS was not to simulate an expert's decision making[1], but to assist the clinician in his or her own decision making.. The first knowledge based systems used were rule based expert systems. It consists of three components –

**Knowledge-base**- It consists of summarised information that is mostly in the form of if–then rules.

**Inference engine** - contains the formulas for combining the rules or associations in the knowledge base with actual patient data.

**A mechanism to communicate with the user-** It is a process of getting the input data, can be a patient data and processing the output from which user can make the accurate decisions.

*Non-knowledge based-*

It makes use of Machine Learning and Statistical Pattern Recognition based methods like Inductive Tree methods, Case based reasoning, Artificial Neural Networks, Genetic Algorithms etc.

Many challenges are also associated with the implementation of CDSS-

A CDSS must be integrated with a health care organization's clinical workflow, which is often already complex. Most clinical decision support systems are standalone products that lack interoperability with reporting and electronic health record (EHR) software. The sheer number of clinical research and medical trials being published on an ongoing basis makes it difficult to incorporate the resulting data. Furthermore, incorporating large amounts of data into existing systems places significant strains on application and infrastructure maintenance.

The very important question that will generally come in one's mind is that when to adopt CDSS for practice. The decision to adopt a CDSS for local patient care is complex and is influenced by many considerations. Those responsible for CDSS implementation are typically administrators, information technology managers, and clinicians, all of whom are increasingly pushed by technology and guided by government regulations. Important issues include CDSS user acceptance,

workflow integration, compatibility with legacy applications, system maturity,and upgrade availability. Some are concerned about increased practitioner dependence on CDSSs, with abraded capacity for independent decision making.

So, Finally, cheaper, non computerized alternatives may be equally or more effective in improving care and reducing medical errors.One of the most important consideration in adopting CDSS is its clinical effectiveness,which is a measure of the extent to which a particular intervention works. The measure on its own is useful, but decisions are enhanced by considering additional factors, such as whether the intervention is appropriate and whether it represents value for money. In the modern health service, clinical practice needs to be refined in the light of emerging evidence of effectiveness but also has to consider aspects of efficiency and safety from the perspective of the individual patient and carers in the wider community.

While some perceive that CDSSs improve efficiency and reduce costs, the current supporting evidence is limited. Although some studies have assessed the costs when outcomes were improved, the cost effectiveness of these systems remains unknown. Many studies suggested the CDSS was inefficient, requiring more time and effort from the user compared with paper-based methods.Finally, most CDSSs used research funding to facilitate implementation.There is currently widespread enthusiasm for introducing electronic medical records, computerized physician order entry systems, and CDSSs into hospitals and outpatient settings. In other commercial, industrial, and scientific fields, computers have become global and have improved safety, productivity, and timeliness. So,because this progress, computerization of the health care environment should offer tremendous benefits. However, multiple challenges have arisen at every phase of software development, testing, and implementation. The progress of CDSSs has mirrored these trends. Systems are proliferating, their technical performance and usability are improving, and the number and quality of evaluations is increasing. These evaluations have shown that many CDSSs improve practitioner performance. However, further research is needed to elucidate the effects
of such systems on patient health.

## 1.5 DECISION SUPPORT FOR PATIENTS

With the rapid growth of computing technology available to consumers and the virtual explosion of health information available on the World Wide Web, patient decision aids and computer-based health interventions are now a more common part of routine medical care.So,with the increase in technology consumers are empowered to take active role in their own health care and to provide the necessary information to enhance their decision making. Research studies have

shown that access to health information can enable patients to be more active participants in the treatment process, leading to better medical outcomes and decision making process. Involvement in one's medical care also involves the concepts of patient empowerment and self-efficacy. Empowerment and self-efficacy are closely related to each other. Empowerment is the process that enables people to "own" their own lives and have control over their destiny. It is closely related to health outcomes. Similarly, self-efficacy is a patient's level of confidence that he or she can perform a specific task or health behaviour in the future. As,there is strong effect of both empowerment and self-efficacy on health outcomes it is very important to keep focus on these concepts while designing the systems for the use of patients.

As medical care increasingly focuses on chronic disease, it is especially important that patient preferences regarding the long-term effects of their medical care be taken into account. For patients to be adequately informed to make decisions regarding their medical care, it is important that they obtain information about the quality of life associated with the possible medical outcomes of these decisions.[7]

Information on patient preferences is important for modifying information to patients and for providing decision support. The modified information has been found to be more effective in providing consumer information and is preferred by patients.In addition to differences in preferences for health outcomes, patients differ in the degree to which they choose to be involved in decision making. Research confirms that age (younger), gender (females), and education level (more) are strong predictors of the desire to be involved in medical decisions.

The computer is used as the health information medium. Computer approaches have the additional advantages of interactivity, providing feedback in the learning process and the ability to tailor information to the individual patient. However, in many cases, more research is required to demonstrate the effectiveness of computer approaches. In addition, designers of systems for patients have not always been sufficiently sensitive to human-computer interface issues.The design of a system for general health education for patients requires specifications that meet a variety of needs. The information and decision aids range from general home healthcare reference information to symptom management and diagnostic decision support.Many diagnostic decision support systems are also developed.

### 1.6 FUTURE OF DECISION SUPPORT SYSTEMS FOR PATIENTS

Advances in communications and information processing technology are certainly changing the way in which medicine is practiced, with dramatic impact on how patients are beginning to receive their health information and interact with the medical care system.[8] There has also been a shift toward consumers becoming empowered participants and assuming a more active role in their

medical care decisions, through increased and more effective access to healthcare information and decision tools. The developers of computer applications for patients have pushed the field of consumer health informatics forward with many innovative systems. However, to achieve significant improvements in the quality of care and health outcomes, researchers and system developers need to focus on bringing the knowledge gained from previous work in health education and behaviour change into the design of new systems. This is a rapidly developing field,with significant innovations in the commercial sector. However, research in several areas is needed to move the field forward in providing real benefits to patients' health outcomes and in showing the effectiveness of the systems to purchasers of health care. The criteria for evaluating computer-based decision support systems for patients are similar to the criteria for physician systems, namely accuracy and effectiveness. However, the rapid deployment of these systems, in an ever changing medical care environment, makes critical evaluation of consumer health information systems extremely difficult.Web sites change daily, and access to one system usually means increased access to many others. It is important to understand the potential effectiveness of investments in this area. Careful needs assessment before system development, usability testing during development, controlled clinical trials, and studies of use and outcomes in natural settings are all critical to our understanding of how to best provide health information and decision assistance to patients.

# 2.DATA WAREHOUSING AND DIMENSION MODELLING

## 2.1 DIMENSION MODELLING

A dimension model is generally the design of data warehouse which helps in storing the information in the form of a model. Dimension models are extensible and can accommodate change. It stores the information in the form of facts and dimensions which can be arranged in star schema or snowflake schema.Facts generally contains all the measurements corresponding to the dimensions, all the measurable values are stored in it.[9] All the content of the data warehouse is stored in the fact table which generally consists of two columns, one storing the facts and the other storing the foreign keys referring to the primary keys in the dimension tables. Different types of measures are stored in it like additive, non-additive etc. Grain is used to define the fact, which is the smallest level by which one is going to define the granularity.

Dimensions are the descriptors which consist of descriptive attributes used to define the fact. All the information of the fact table is stored in the dimensions. Dimension table attributes play a very important role in making the data warehouse understandable and usable. Dimensions used to store primary keys which correspond to the foreign keys in the fact table. It has been observed that more the quality of the values in the attribute columns more better is the data warehouse.So,the dimension attributes are directly proportional to the quality of the data warehouse to be developed.

The dimensions whose attributes changes over a period of time are known as slowly changing dimensions(SCD).These are basically of three types which are of type 1,type2 and type 3,where type 1 is used when there is no need to take into account historical changes,type 2 when it is necessary to take into account the historical changes and the last one type 3 which is used when it is necessary to keep record of historical changes which only occurs for a finite time period.
Data modeling is building of the data models which shows relationships between data.
There are three types of data model which are conceptual, logical and physical.

In the conceptual data model relationship between different entities is shown, in logical data model the details of the data is considered without taking into consideration the implementation whereas finally the physical data model where the exact implementation of the data model in the database is known. It is the physical data model where entities are converted to tables, relationships to foreign keys and attributes to columns.

But the data warehouses are generally developed using dimension models as they are very flexible for the user perspective and contains denormalized data whereas data models are not flexible and contains normalized form of data.

Normalization of data is done to remove the redundancy and dependency of the data.

Denormalized data is used in data warehouse data modeling because redundancy is required to provide an efficient, high-qualified and extendable decision making support to high level managers or decision makers.

## 2.2 DATA WAREHOUSING

A data warehouse is a database used for storing current and historical data and integrating the data from various operational sources where the data is stored in various file formats.It is a database mainly used for reporting and the analysis of data.

A data warehouse is an isolated database which is used across an enterprise to combine data from different data stores and serve all business task supporting systems with a unified view of business data**.** A data warehouse maintains its functions in three layers: staging, integration, and access.[10]They were developed to meet a growing demand for management information and analysis that could not be met by operational systems.
Components of a Data Warehouse include Operational Source Systems, Data

Staging Area, Data Presentation Area, and Data Access Tools. Data warehouse is characterized by a strict separation of operational and decisions-making data and systems. It is a place where data is stored for archival, analysis and security purposes. Usually a data warehouse is either a single computer or many computers (servers) tied together to create one giant computer system. Effective data warehouse had to be integrated, subject oriented, nonvolatile, and time

variant in nature A data warehouse is sometimes said to be a major role player in a decision support system (DSS). DSS is a technique used by organizations to come up with facts, trends or relationships that can help them make effective decisions or create effective strategies to accomplish their organizational goals.Data integration allows us to assemble targeted data reagents for bioinformatics analyses, and to discover scientific relationships between data.
Integrating these disparate sources of data enables researchers to discover new associations between the data, or validate existing hypotheses. Data can consist of raw or formatted form. Data is most valuable of all these and good data should fulfill at least these:-

(1) Data has to be accessible,

(2) Data has to be current,

(3) Data has to be flexible, and

(4) Data has to be understandable

Some of the benefits that a data warehouse provides are as follows:

- · A data warehouse provides a common data model for all data of interest regardless of the data's source. This makes it easier to report and analyze information than it would be if multiple data models were used to retrieve information such as sales invoices, order receipts, general ledger charges, etc.

- · Prior to loading data into the data warehouse, inconsistencies are identified and resolved. This greatly simplifies reporting and analysis.
- · Information in the data warehouse is under the control of data warehouse users so that, even if the source system data are purged over time, the information in the warehouse can be stored safely for extended periods of time.
- · Because they are separate from operational systems, data warehouses provide retrieval of data without slowing down operational systems.
- · Data warehouses can work in conjunction with and, hence, enhance the value of operational business applications, notably customer relationship management (CRM) systems.
- · Data warehouses facilitate decision support system applications such as trend reports (e.g., the items with the most sales in a particular area within the last two years), exception reports, and reports that show actual performance versus goals.

- · Data warehouses can record historical information for data source tables that are not set up to save an update history.

**Components of data warehouse-**

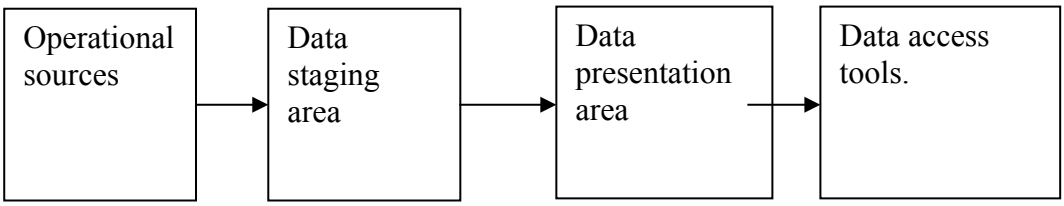



**Figure1: Elements of data warehouse**

First the data is collected from various operational sources which stores data in various file formats. The data staging area of the data warehouse is used for storage of the data and a set of processes which are referred to as extract-transform-load (ETL).

The data staging area is the area between the operational source systems and the data presentation area. As, it can be seen that ETL i.e extraction ,transformation and loading is followed in data warehouse building . Extraction is the first step in the process of getting data into the data warehouse environment. Extracting the data comprises of understanding the source data and then loading the data which is needed for the data warehouse into the staging area for various other transformations and changes.

After the extraction of the data into the staging area various transformations like cleaning ,combination of data from various sources etc .All these transformations are applied before the data is loaded to the presentation layer. Clean up and transformation tools help in it. These tools also help in the maintenance of meta data.

Meta data is nothing but the data within the data which helps in describing the data warehouse more efficiently. It helps in the construction, maintenance and the management of the data warehouse. Meta data makes the data easily accessible to the users and thus it will help in better understanding the content of the data. These clean up tools have to come across with some of the major issues like database heterogeneity and data heterogeneity. These tools are necessary as they help in removing the unwanted data from the operational database ,forming the default values for the missing values etc.

In the presentation layer there is an organization and storage of data so that users can query directly. The last and the most important component of the data warehouse is the data access tools. The main aim of data warehouse is to provide all the information to the users for the decision making purpose and this is done with the help of various access tools which can be reporting tools, online analytical processing tools,data mining tools etc. where OLAP tools are based on dimensional data models and they help in analysing the data using multidimensional views.

So, data warehouse is a very important concept the field of data management system.

Main goals of data warehouse are as follows-

- It helps in making the information easily accessible to the user
- It is dynamic in nature and adaptive to the change
- It helps in making the improved decision making.

## 3. KNOWLEDGE DISCOVERY AND DATA MINING

Data mining is the process of identifying the hidden knowledge in the large sets of data,so it discovers the hidden knowledge which is helpful in making further decision making processes. The main goal of data mining is to discover new patterns for the user where these patterns help in description and prediction purposes which means finding the patterns and presenting them to the users in an understandable form and to identify those variables which help in predicting the future values.[6]

Data mining is used to provide decision support in the healthcare setting. There is lot of pressure in the healthcare organizations to improve the quality of care without increasing the costs. So, this data mining helps in enhancing practices of the physicians, disease management etc.

Knowledge discovery data mining which is an iterative process consists of following steps-

- The business understanding
- Selection of data set,
- Cleaning and preprocessing of data
- Data reduction and projection,
- Matching of the objective defined in step 1 into a data mining method (classification, clustering, regression, etc.),
- Choice of the algorithm and search for various data patterns,
- Extraction of the hidden patterns
- Interpretation of the data and use of the knowledge discovered.

## 3.1 DATA MINING AND PATTERN RECOGNITION

One of the major tasks in data mining is the pattern recognition .It helps in taking the information from the models and describes and classifies the measurements associated with it.The main aim is to get the knowledge from this and try to facilitate decision making processs further. Many approaches are there in pattern recognition of which one of the approaches commonly in use is statistical pattern recognition.It uses statistical approach to modelling of the measurements .The main aim is to select that feature which helps in classification of the objects.[8]

With the advancement in the medical technology more and more methods for data classification have been developed due to which data mining has got many applications in clinical decision support systems.There are two types of decision support systems i.e which employ or consider data mining tools and the other ones that consider rule based expert systems.Decision support systems which employ rule based expert systems require large amount prior knowledge so that the decision maker can form the right decision whereas in case of decision support system employing data mining tools do not require any king of prior knowledge as it helps to find the unusual ,unexpected ,hidden patterns or relationships in the data,then the system will apply this newly discovered knowledge to the newly formed data set.

Generally data mining has 3 categories-

- *Supervised learning*
- *Unsupervised learning*
- *Semi-supervised learning*

*Supervised learning-*

In case of this type of learning the user beforehand knows what type of classes are there and then transfers this knowledge to the training process which consists of training data sets having dependent or independent variables. The system is adjusted based on the difference between the desired result and the actual result. The main aim of this type of learning is to reduce the disparity between the expected and the observed result. Supervised learning helps in construction of predictive model which will predict the future values or the behaviour of an entity. Here, a target variable is there which will determine the functionality of other variables.

*Unsupervised learning-*

In case of this type of learning no prior knowledge is there about how to classify the groups into meaningful classes. Classification of groups occur because of the similarities identified by the learning system. All the variables are considered similar without any distinction to dependent and independent variables. Target variable is also absent.

*Semi-supervised learning-*

In this type of learning, the target variable is present but consists of small amount of labelled data and a lage amount of unlabelled data.This type of learning falls between supervised and unsupervised learning.So,the data mining techniques help in estimating the values of the missing target values or to extract hidden patterns,relationships or clusters in the given data set.

## 3.2 CLINICAL DATA MINING

Clinical data mining deals with the applications of the data mining to solve the clinical problems. First step is to understand the clinical data properly.Clinical information systems manage the electronic health record of each patient.Clinical data warehouse is created for the better understanding,care and the effectiveness. Data mining helps in exploring the hidden information stored in the data.The approaches used in this process are data visualization,data exploration and data quality assessment. Data visualization approach helps in converting the data to information.Data exploration provides with the relationships in the clinical data which are generally expressed in the form of If conditions and Then conclusions.

The clinical data contains noise,missing values and sometimes the unstructured data. Missing values generally arise because of the neglection of the clinicians as they thought that the some variables are of less importance for a specific patient. In the case of data quality assessment it access the data thoroughly and helps in detection of medical errors,checking the data coding quality and providing a structure from unstructured data etc.

There are many applications of the data mining in the clinical field . It may help in providing the diagnostic support for the events which occur rarely like some diseases which are very rare ,eg, otoneurological diseases and can also be helpful in difficult diagnosis. The improvement of the quality of care concerns mainly support for patient safety.It can also help in improvement of the administrative work which is done by the healthcare staff. In order to handle the clinical data in a better way data mining techniques have evolved some solutions which will help in classifying the medical reports automatically by preserving the privacy of the patient successfully.

# CHAPTER 2

# LITERATURE REVIEW/BACKGROUND

Bioinformatics have evolved a great use in the clinical or medical field. Many techniques have been used in the this field in order to solve complex problems related to the medical field due to which many patients and staff suffers. Because of the complex problems decision making process has become difficult. Various type of Decision support systems have been developed in the past years according to the various problems which have arisen in the medical system.

One such type of expert medical system has been developed by M.J. Sawar, T.G. Brennan,* A.J. Cole, J. Stewart*(1992) POEMS i.e Post operative expert medical system which is a decision support system developed to manage complications or the problems like when to take specific action based on monitoring signs which occur during the post operative care.

Post operative care Postoperative care refers to the care one receive following a surgical procedure. This may include pain management and wound care. The type of postoperative care one require depends on the type of surgery one has[2].

.POEMS interactively receives data obtained from the patients based on the standard strategy used by the medical staff: History, Examination, and Investigative tests.

When asked for diagnosis it presents an ordered list of likely, possible and not-likely candidate diagnoses. [3]It can then answer questions on how the diagnosis was reached, what treatment would be most suitable, or what further investigative action could be taken to focus on a particular diagnostic candidate. The patient data is maintained as patient history and can be used later to generate trends and other time dependent diagnoses.

The medical knowledge in POEMS is divided into three categories which are generally known as KR objects-

(i) Medical domain knowledge, like blood pressure, haemoglobin etc.

(ii) Candidate diagnoses knowledge, like congestive heart failure etc.

(iii) Treatment actions knowledge, like antibiotics, analgesics etc.

These categories are connected by various links like supporting ,action and caution link respectively.

Data for the patient is entered into POEMS by pressing new patient button and the user can also enter data collected by taking history, making examinations and performing various investigative test. [4]All these data collection classes represented as buttons which can be pressed to bring up menus for the individual operation (Myers B A. *The Garnet Toolkit Reference Manuals*, CMU-CS-89-196, Carnegie Mellon University, March 1990.)

Graphical representation of the diagnostic process is explained where all the data objects for a particular patient entered into POEMS through data acquisition stage , with their links in place.

The diagnoses candidates with their discriminatory, strong support, supporting and exclusionary slots are also representated which determine the likelihood of the candidates with links attached to data objects on one side and their suitable treatment actions on the other side. The diagnosis is carried out incrementally as the user enters the patient data. It is based on the constraint propagation mechanism of KR, which automatically manages all the constraints and maintains consistency.[11] This is a distributed mechanism which is better than the diagnostic process based on ATMS which is described by Sawar M J et al. in (1991)Brennan T G, Cole A J and Stewart J. POEMS (PostOperative Expert Medical System), in *Proceedings of IJCAI-91 one Day Workshop : "Representing Knowledge in Medical Decision Support Systems"*,Sydney, Australia, Aug. 1991.). The ATMS diagnoser was initiated by collecting all the medical domain knowledge data objects acquired for the patient, and passing them to the justification algorithms as explained by(Kelleher G. A brief Introduction to Reason Maintenance Systems. in *Reason Maintenance Systems and their Applications*, ed. Smith and Kelleher pp. 4-20, U.K., Ellis Horwood.).

Once the candidate diagnoses are reported, the user can query the various aspects of the diagnoses, where a menu of diagnosis related questions is displayed.

The user can also query the treatment actions which are appropriate for a selected candidate. The knowledge acquisition used two learning methodologies i.e Active and Passive learning. In Active learning the system queries the expert about the validity of new diagnostic candidates, and whether they are likely, possible or not likely(Sawar M J and Thomas R C. Learning Apprentice System for Turbine Modelling. in *Proceedings of IEA/AIE-90*, Charleston, South Carolina, U.S.A. July 1990.)whereas in passive learning system acquires individual medical domain knowledge without any queries. This, generates a big disadvantage of using passive learning methodology as expert is never in the position to know exactly what has been learnt from the scenario presented to the system, and what more needs to be taught in order to fill the gaps left by the previous scenarios.

The postoperative healthcare team is under constant pressure to discharge patients quickly. This can lead to vital signs being missed and result in a delay in recovery.

Patients can be discharged quickly only when they do not experience any post-operative complications, many of which can be avoided or identified with correct and thorough monitoring of signs and symptoms, diagnosis and by following proper treatment actions and other investigative actions which can be successfully achieved with the help of POEMS.

Similar concept can be applied in order to reduce the time taken in the hospitals to go through the clinical sets and determine the critical factors which increases the time factor and thus extending the time of the patients.Depending on the observations observed the disease can be determined and the likelihood of the disease can also be determined and depending on the disease various investigative tests can be applied further which will help in reducing the time and decreases the complexity of the medical records.

According to Jonathan C. Prather, M.S. et.al(1994) medical data mining techniques also known as knowledge discovery techniques in databases have been applied in order to discover hidden patterns responsible for discovering new medical knowledge.

Large quantities of data are generated through the health care process. Many computer based technologies like computer based record software helps in making the data easily accessible and manageable but in order to analyze and evaluate the data few tools exist which help in discovering the hidden patterns or trends which can increase the understanding of the clinical data,thus making the knowledge about the disease progression more enhanced and manageable. [12]Techniques are needed to search large quantities of clinical data for these patterns and relationships. Past efforts in this area have been limited primarily to epidemiological studies on administrative and claims databases. These data sources lack the richness of information that is available in databases comprised of actual clinical data.

Data mining, also referred to as Knowledge Discovery in Databases or KDD, is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amounts of data as mentioned earlier also. The typical data mining process involves transferring data originally collected in production systems into a data warehouse, cleaning or scrubbing the data to remove errors and check for consistency of formats, and then searching the data using statistical queries, neural networks, or other machine learning methods . Most previous applications of KDD have focused on discovering novel data patterns to solve business related problems such as designing investment strategies or developing marketing campaigns. Data warehousing and mining techniques have rarely been applied to health care[13]. ,Researchers at the Southern California Spinal Disorders Hospital in Los Angeles used data mining to discover subtle factors affecting the success and failure of back surgery which led to improvements in care . In a second health care application, GTE Laboratories built a large data mining system that evaluated

health-care utilization to identify intervention strategies that were likely to cut costs.This system, however, is focused on cost analysis and not on identifying new associations or relationships within clinical data.

Production System Database used for mining comprised of The Medical record or known as TMR,where TMR is a computer based patient record system developed at Duke University over the last 25 years. This TMR comprises of the problems, therapies, summaries etc and stores all the information of the patient in a single record. The database selected was perinatal which is used by by the Department of Obstetrics and Gynecology at Duke University Medical Center.

Next step followed is Extracting and Cleaning the Dataset for Analysis.For the purposes of this study, a sample two-year dataset (1993-1994) from the data warehouse was created to be mined for knowledge discovery.

Multiple SQL queries were run, data was cleansed of error values, missing values etc which was done by Paradox application language scripts ,where the main challenge of this is to convert alphanumeric fields into numeric values .

Then lastly in Mining the Dataset exploratory factor analysis technique was applied which identifies the data elements used to explain the differences between different patient groups.This technique is applied where there are large number of subjects being compared on a set of rules.

Some descriptive models have been also discovered in medical care by G.Stiglic, P.Kokol. Models have been examined for medical housestaff,pharmacy services,and social workers. They have been considered for ambulatory care, home care, and nursing homes. Care models also exist for specific patient populations such as elderly patients,people with mental health needs, and individuals with chronic conditions to include disease management models and the use of technology.

Chronic disease model as described by Chirch LM ,Hasham M, Kuchel GA.in 2013 which helps in the management of chronic diseases. With the discovery and widespread use of antiretroviral therapies, [14.15]now more and more numbers of individuals with HIV are now able to live into advanced age without any fear. So,individuals involved in HIV care, policy, and research have increasingly had to refocus their efforts from a traditional infectious disease emphasis toward conceptual models grounded in the management of common chronic diseases and geriatric syndromes.[16]These conceptual models help in giving the proper care to the patients suffering from AIDS and also improving the quality of care.

Schweitzer M et.al described the clinical workflows for diabetes care.Basically in this the ontohealth project was described which enhances the use of EHRs to a more comprehensive

integration.As, Electronic health records (EHRs) play an important role in the treatment of chronic diseases such as diabetes mellitus. The purpose of the OntoHealth project is to foster a functionally flexible, standards-based use of EHRs to support clinical routine task execution by means of workflow patterns and to shift the present EHR usage to a more comprehensive integration concerning complete clinical workflows.A categorization model was developed which allows for a description of the components or building blocks of clinical workflows from a functional view.These models developed are generally related to some specific diseases,models have also been developed which are independent of the diseases but are very rare.

Next,a clinical staging model was also developed to provide quality of care to the youth suffering from mental health problems in Australia. Cross SP, Hermens DF,et.al proposed propose a clinical staging model that has the potential to better match illness stage to intervention.
The model allows clinicians to provide more personalized and responsive care, especially to young people with attenuated syndromes who have a clear need for mental health care but who may not otherwise receive it. This approach can also assist clinicians in considering the potential trajectory of illness.

So,as we can see from the past researches that this clinical informatics field has evolved much by developing the models which are helpful in improving the quality of care and safety of the patient.

Much work has also been found on integrating the large scale clinical data to improve the clinical care. Electronic medical records have gained much importance. In clinical informatics, the widespread adoption of the EMR system has generated large amounts of heterogeneous clinical data-some structured and others unstructured.[17] In genetics, since the completion of the Human Genome Project in 2003, the acquisition, analysis, and presentation of whole-genomic data has become faster, cheaper, and more reliable day by day . Such dramatic technological advances affect the development of new prevention, diagnosis, and treatment patterns for routine clinical care. The massive amount of heterogeneous data from two different domains is expected to provide personalized, preventive, and predictive healthcare services in the near future.

Integrated use of EMR and bioinformatics is beginning to influence the changes in the research paradigm-that is, rapid introduction of new concepts into the point of care. Dr. Want used clinical bioinformatics (CBI) with the definition of "the clinical application of bioinformatics-associated sciences and technologies to understand molecular mechanisms and potential therapies for human disease" CBI aims to deal with the challenge of integrating genomic and clinical data to accelerate the translation of knowledge into effective treatment plan development and personalized prescription.[18,19] It is to assist clinicians in various ways, including new biomarker discovery,

identification of genotype and phenotype correlations, and pharmacogenomics at the point of care. Biomedical informatics is defined as an emerging, multi-disciplinary field, and it is the integration of the computational methods and diverse technologies used in life science research, such as genomics, proteomics, systems biology, computer sciences, and healthcare applications, such as electronic health records (EHRs).

Genome enabled EMR also has some challenges associated with it like integration of heterogeneous data into one database system. It is not possible to move large amounts of genomic data. The integrated database can have a potential impact on the prevention, diagnosis, and treatment of disease. To make this desire come true, it is important to connect genomics data with clinical information. Genetic test results are already used to assess the risk of breast cancer patients, determine the potential adverse drug reactions on individual patient metabolism, and identify treatment plans for cancers. The integrated use of EMR and BI data needs to consider four key informatics areas: data modeling, analytics, standardization, and privacy.

Genomic technologies hold the potential to improve the diagnosis and treatment of inherited and complex diseases-including cancer- and facilitate the move towards personalized predictive medicine. The higher throughput and rapidly falling costs of next-generation sequencing have resulted in voluminous genomic data and downstream computational challenges. Thus, the shift from this powerful discovery research to clinical implementation can only be accomplished with careful integration with EMRs, a frontline patient care tool[20,21,22]. The most prominent reason to integrate clinical information and biology information under the same system is to provide opportunities for bi-directional exchange of data, technology, and knowledge between two disciplines with different histories and cultures. Additionally, open global cooperation will provide opportunities to make rapid progress in understanding, treating, and preventing human diseases.

Many researchers have researched a lot on improving the quality of care for the patients and improving the clinical care by number of ways but mostly focusing on the particular disease,not much work has been done on reducing the time taken in managing the clinical records for the care of the patient and reducing the complexity of the records.So,our main focus is to identify the critical factors and reduce the time taken for the safety and care of the patients. Much of the work has also been found in the field of data mining which helps in finding the hidden patterns in large clinical data sets .

The term “clinical data mining” indicates the data mining application on clinical problems. Understanding the clinical data is very important. Data visualization approaches tend to provide quick and understandable access to information i.e. converts data into information. Visualizing time-oriented data of a patient population is a challenging task and it was proposed by Klimov and Shahar in . While most of data mining applications analyze retrospective data, Chen et

al. proposed a real-time data summarization by parsing hospital communication messages .In data accessing Jannin and Morandi, for example, use data mining to assess the quality of a surgical procedure model . Spangler et al. investigated the adequacy and effectiveness of two coding classifications (ICD-9 for diagnostic and CPT for procedure) in two hospitals . Chapman et al. proposed a method for extracting a clinical concept from emergency department reports and Goldstein et al. classified automatically radiology reports with ICD-9- CM classification.

For assistance in clinical care many expert systems have been developed. In the 70s and 80s, many expert systems such as MYCIN or INTERNIST were developed based on knowledge provided by medical experts.But the acquisition of the knowledge used by these systems has a high cost and artificial intelligence was used to extract knowledge in the clinical data. These systems were designed for one of the following purposes: prognostic assistance, diagnostic assistance, quality of care improvement, and support to access clinical data.[23,24] Gellerstedt et al., for example, provided a support for pre-hospitalized patients suffering from acute myocardial infarction at the emergency department dispatch center using subjective information provided by phone and Goletsis et al. proposed a system for early detection of high risk patients suffering from myocardial ischemia using electrocardiographic data.Daemen et al. compared the breast cancer prognosis model provided by clinical versus genetic data.

For the data mining many algorithms have been developed in the past.For example Harper compared different classif ication algorithms for decision making in (Harper PR. A review and comparison of classification algorithms for medical decision making. Health Policy 2005;71:315-31.)and concluded that there was no single best classification tool and the best performing algorithm not only depends upon the features but also how easily accessible it is to the user.Because of the complexity of medical data, it is sometimes necessary to adapt existing algorithms or optimize their use to obtain better results.[25] Hripcsak et al., for example, proposed and evaluated a new distance metrics for narrative clinical data for clustering Juhola and Laurikkala proposed a new distance measure in the case of mixedtype variables (quantitative and nominal) for classification [Juhola M, Laurikkala J. On distance computation in space of mixed-type variables in medical data mining. Stud Health Technol Inform 2002;90:425-30]. Ramoni and Sebastianini proposed a new version of the Naïve Bayes classifier to handle missing values automatically [Ramoni M, Sebastiani P. Robust Bayes classifiers. Artificial Intelligence 2001;125(1-2):209-26.]. In essence, data mining exploits a large number of variables and measurements. Computational efficiency and scalability are important issues. To address this problem, Huang et al. investigated a new feature selection algorithm to reduce the computational complexity of data mining (Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med 2007;41:251-62.). The heterogeneity of the medical data

prompts medical data miners to develop new approaches to analyze data. Jesneck et al., for example, investigated decision fusion as a strategy for the classification of imaging data from multiple modalities, multiple sources and having various types of features (Jesneck JL, Nolte LW, Baker JA, Floyd CE, Lo JY. Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis. Med Phys 2006;33:2945-54.). The analysis relationships of time-stamped or time series clinical data exploits the temporal abstraction mechanism (identification of time interval in which a specific data pattern occurs) The introduction of a knowledge base represented in an ontology

was introduced by( Tusch G, Bretl CE, Connor M, Das A. SPOT Towards Temporal Data Mining in Medicine and Bioinformatics. In: AMIA Annu Symp Proc 2008. p. 1157. and Raj R, O'Connor MJ, Das AK. An ontology-driven method for hierarchical mining of temporal patterns:application to HIV drug resistance research.[26] AMIA Annu Symp Proc 2007;:614-9) in order to improve the mining of temporal associations in clinical data.This was the work done on data mining methodology development process.

Next step was creation of the structured data set. This can be performed by consulting medical knowledge sources to identify relevant variables for the analysis or with the objective help of experts of the domain. Two studies (Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput Biol Med 2006;36:1351-77., Goodwin LK, Prather JC. Protecting patient privacy in clinical data mining. J Healthc Inf Manag 2002;16:62-67.) analyze the whole clinical data warehouse for exploration and to extract new knowledge inherent in the data. Pregnant women data collected from a 20 year-long perinatal database was used in (Goodwin LK, Prather JC. Protecting patient privacy in clinical data mining. J Healthc Inf Manag 2002;16:62-67) and the whole clinical data of 13 years were analyzed in (Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput Biol Med 2006;36:1351-77). Creating a target dataset may be time consuming especially if the important variables to study have to be inferred from multiple variables from multiple sources[27] . Patient privacy and confidentiality is also an important issue when the data are used out of their initial collection purpose.

Some research use publicly available datasets as those provided by universities like the UCI6 repository or those provided by learned societies such as the Mammographic Image Analysis Society or ImageCLEF7 . These publicly available datasets serve as benchmarks to evaluate the performance of newly developed algorithms.

Care models have also been developed in the field of nursing care. Despite the interest in a variety of care models, it is difficult to discern which models work best. Neither the traditional nor

the nontraditional inpatient nursing care models have been evaluated rigorously for their effects on patient safety. Emerging models from other care disciplines, other settings, and particular patient populations are also lacking rigorous empirical assessments of their relationship to patient safety.

A number of investigations examining care models addressed nurses' perceptions of the care model. Only two investigations combined the nurses' perceptions with patient safety measures. [28] Ngersoll and Redman and Jones were among the first investigators to assess the effects of patient-centered care models on nurse managers. The data from both of these studies expose the pressure and role confusion experienced by nurse managers. Subsequently, a quantitative investigation found nurse managers experienced a high level of emotional exhaustion, a key component of burnout.

Very little is known about the relationship between the care models and safety of the patients. Care delivery models range from traditional forms, such as team and primary nursing, to emerging models. Even models with the same name may be operationalized in very different ways. The rationale for selecting different care models ranges from economic considerations to the availability of staff. What is glaring in its absence, however, is the limited research related to care models. Even more sparse is research that examines the relationship between models of care and patient safety. Ideally, future studies will not only fill this void, but the models tested will be developed based on a comprehensive view of patient needs, taking the full complement of individuals required to render quality care into account.

# CHAPTER-3

## CONTRIBUTIONAL WORK, DESCRIPTION AND RESULTS.

The problem related to clinical datasets is that they are very large ,so are very time consuming. All the data contains some hidden knowledge ,patterns or relationships which will help in understanding of the data properly and help in better decision making process. In order to reduce the time taken in clinical datasets these hidden patterns which are critical factors which play a main role in increasing the time factor in these large clinical data sets are identified which will reduce the complexity of the medical records and will make them easily accessible and understandable for the user.

The information is mined and extracted from the model which is developed for the patient care which will help in improving the safety and quality of care for the patients.

**Tools and technologies used-**

*Erwin data modeller-*

It is tool for creation of data models.When you open it and go to the file menu and then to new option it asks for what type of model you want to create whether logical,physical,logical/physical. The basic and the most important difference between logical and physical model is that physical model is platform independent ant the entities which are in logical model are converted to tables in physical model and attributes are converted to columns in the physical model.

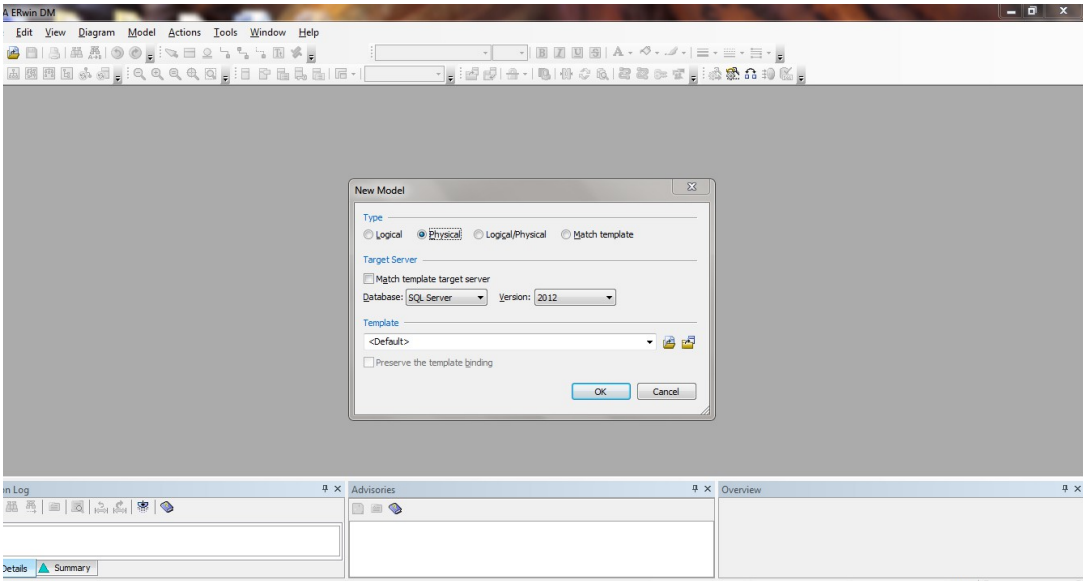



**Figure2: Type of model creation in Erwin Data modeller**

On selecting the required option,like in my case I have selected physical the following screen appears, where then dimensional model can be made by choosing the options of tables and relationships.

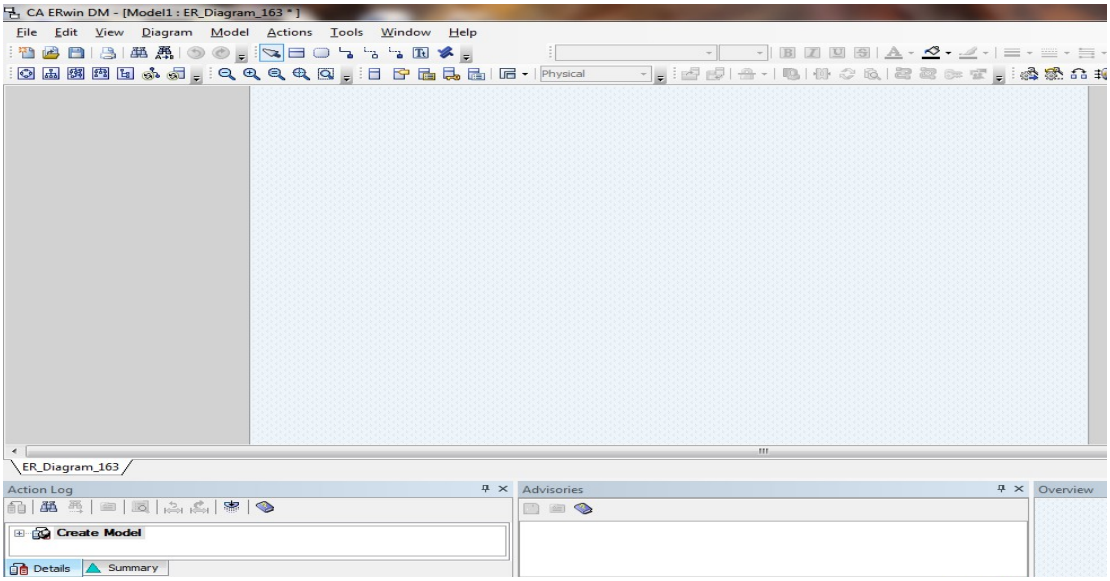



**Figure3: Erwin Data modeller**

*Mysql-*

It is a relational database management system(RDBMS) developed after the name of daughter of the developer Michael Widenius The SQL stands Structured Query. It includes following features-

- A broad subset of ANSI SQL 99, as well as extensions
- Cross-platform support
- Stored procedures
- Triggers
- Cursors
- Updatable Views
- True Varchar support
- Information schema
- Strict mode
- X/Open XA distributed transaction processing (DTP) support; two phase commit as part of this, using Oracle's InnoDB engine
- Independent storage engines (MyISAM for read speed, InnoDB for transactions and referential integrity, MySQL Archive for storing historical data in little space)
- SSL support

- Query caching
- Sub-SELECTs (i.e. nested SELECTs)
- Replication support (i.e. Master-Master Replication & Master-Slave Replication) with one master per slave, many slaves per master, no automatic support for multiple masters per slave.
- Full-text indexing and searching using MyISAM engine
- Embedded database library
- Partial Unicode support (UTF-8 and UCS-2 encoded strings are limited to the BMP)
- Partititoned tables with pruning of partitions in optimizer
- Shared-nothing clustering through MySQL Cluster
- Hot backup (via mysqlhotcopy) under certain conditions.

In this query is written in the query box and executed. Database is created both for functional as well as staging schema.First the staging database is created named as Staging_project under which different tables are created in order to dump or load,store the data in these tables from the csv files already prepared. After this tables under the functional schema database is created under which dimension and fact table is created.

*Pentaho-*

It is an open source Business Intelligence software with integrated reporting, dashboard, data mining, workflow,Data Warehousing and ETLcapabilities. Its headquarter is in Orlando, USA. It offers several products such as Pentaho Data Integration, Pentaho Analysis Services, Pentaho Reporting, Pentaho Data Mining, Pentaho DashBoard, Pentaho for ApacheHadoop.

- Pentaho Data Integration- Data Integration in pentaho is done by kettle. It consists of a core data integration engine, and GUI applications that allow the user to define data integration jobs and transformation.

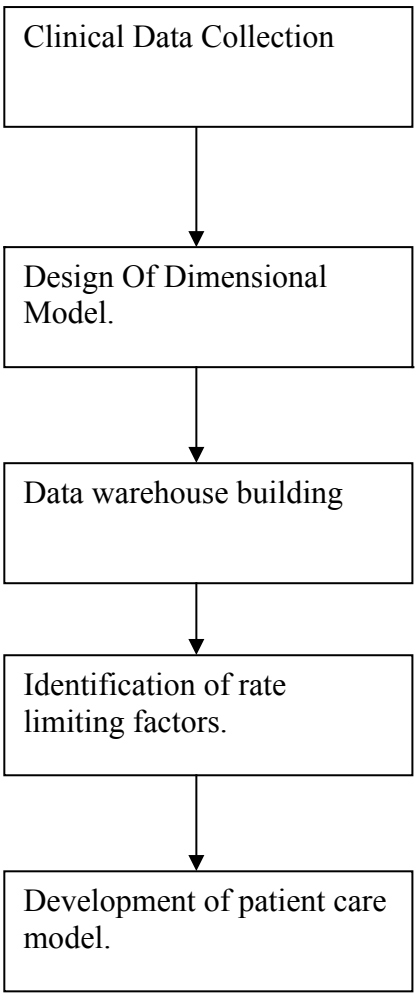
Pentaho Data Integration (PDI) is a flexible tool that allows you to collect data from different sources such as databases, files, and applications, and turn the data into a unified format that is accessible and relevant to end users.

Common Uses of Pentaho Data Integration Include:

- Data migration between different databases and applications.
- Loading huge data sets into databases taking full advantage of cloud, clustered and massively parallel processing environments
- Data Cleansing with steps ranging from very simple to very complex transformations

- Data Integration including the ability to leverage real-time ETL as a data source for Pentaho Reporting
- Data warehouse population with built-in support for slowly changing dimensions and surrogate key creation

**FLOWCHART OF THE PROCESS-**




Clinical data collection is done by collecting the data from the record section of the IGMC,Shimla.

*Design of dimensional model-*

In Erwin data modeller-Dimensional model is created for the patient care in Erwin data modeller.Physical dimensional model is created which consists of foreign keys,primary keys,tables ,columns and is platform independent.

Model created is as follows-

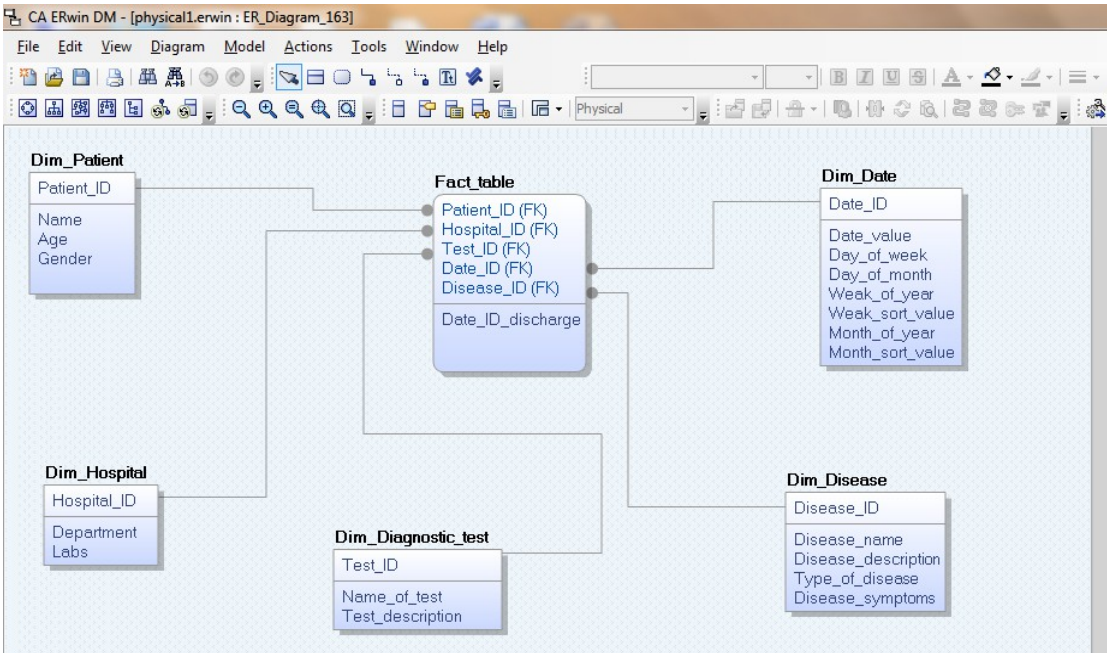



**Figure 4: Physical Data Model**

*Construction of database in mysql-*

Functional database is created in mysql version1.1.20 named as functional_project and by using the syntax as mentioned in appendix.

*Design of data warehouse –*

The construction of data warehouses, which involves data cleaning and data integration, can be viewed as an important preprocessing step for data mining. Moreover, data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Furthermore, many other data mining functions such as classification,prediction, association, and clustering, can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction. Hence,data warehouse has become an increasingly important platform for data analysis and online analytical processing and will provide an effective platform for data mining.

The construction of a data warehouse requires data integration, data cleaning, and data consolidation. The utilization of a data warehouse often necessitates a collection of decision support technologies. This allows knowledge workers (e.g., managers, analysts, and executives) to use the warehouse to quickly and conveniently obtain an overview of the data, and to make sound decisions

based on information in the warehouse. Some authors use the term \data warehousing" to refer only to the process of data warehouse construction, while the term warehouse DBMS is used to refer to the management and utilization of data warehouses.

Data warehousing is also very useful from the point of view of heterogeneous database integration. Many organizations typically collect diverse kinds of data and maintain large databases from multiple, heterogeneous, autonomous, and distributed information sources. To integrate such data, and provide easy and efficient access to it is highly desirable, yet challenging. Much effort has been spent in the database industry and research community towards achieving this goal.

In this step, we'll populate a data warehouse with data from the OLTP system. This phase of the process is known as ETL, which stands for Extract, Transform, Load. This is exactly what needs to be done. Extract the data needed for the fact and dimension tables from all different data sources, transform it to fit our needs and load it into the data warehouse so it can be queried. Some important terms before we go with construction of Data warehouse are as follows-

- Metadata- It is referred as "data about data". Metadata is all the information in the data warehouse environment that is not the actual data itself. Metadata is a loose term for any form of auxiliary data that is maintained by an application. Metadata is also kept by the aggregate navigator and by front-end query tools. The data warehouse team should carefully document all forms of metadata. Ideally, front-end tools should provide for tools for metadata administration. Most of the extraction steps should be handled on the legacy system. This will allow for the biggest reduction in data volumes. is structured data which describes the characteristics of a resource. It shares many similar characteristics to the cataloguing that takes place in libraries, museums and archives. The term"meta" derives from the Greek word denoting a nature of a higher order or more fundamental kind. A metadata record consists of a number of predefined elements representing specific attributes of a resource, and each element can have one or more values.

  Each metadata schema will usually have the following characteristics:

  - a limited number of elements
  - the name of each element
  - the meaning of each element
- Data mart- A data mart (DM) is the access layer of the data warehouse (DW) environment that is used to get data out to the users. The DM is a subset of the DW, usually oriented to a specific business line or team. A data mart is a data repository that may or may not derive from a data warehouse and that emphasizes ease of access and usability for a particular designed purpose. There can be multiple data marts inside a single corporation; each one relevant to one or more business units for which it was designed. DMs may or may not be dependent or related to other data marts in a single corporation. If the data marts are designed using conformed facts and dimensions, then they

will be related. In some deployments, each department or business unit is considered the *owner* of its data mart including all the *hardware*, *software* and *data*. A database, or collection of databases, designed to help managers make strategic decisions about their business. Whereas a data warehouse combines databases across an entire enterprise, data marts are usually smaller and focus on a particular subject or department. Some data marts, called *dependent data marts,* are subsets of larger data warehouses.

- Data Normalization- Data Normalization means to bring down data into same level so that some conclusion can be formed from it. In the design of a relational database management system (RDBMS), the process of organizing data to minimize redundancy is called normalization. The goal of database normalization is to decompose relations with anomalies in order to produce smaller, well-structured relations. Normalization usually involves dividing large tables into smaller (and less redundant) tables and defining relationships between them. The objective is to isolate data so that additions, deletions, and modifications of a field can be made in just one table and then propagated through the rest of the database via the defined relationships. Edgar F. Codd, the inventor of the relational model, introduced the concept of normalization and what we now know as the First Normal Form (1NF), Second Normal Form (2NF) and Third Normal Form (3NF) in 1971, and Codd and Raymond F. Boyce defined the Boyce-Codd Normal Form (BCNF) in 1974. The higher the normal form applicable to a table, the less vulnerable it is to inconsistencies and anomalies. Each table has a "highest normal form" (HNF) by definition, a table always meets the requirements of its HNF and of all normal forms lower than its HNF; also by definition, a table fails to meet the requirements of any normal form higher than its HNF.

- Data cleaning- Data cleaning is getting data into consistent form. Data cleansing or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table,or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant etc. parts of the data and then replacing, modifying or deleting this *dirty data*.After cleansing, a data
  set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by different data dictionary definitions of similar entities in different stores,may have been caused by user entry errors, or may have been corrupted in transmission or storage. The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. A two step process including the detection and the correction of the errors.

*Construction of data warehouse*

When we move the data into data warehouse normally, it will have two kinds of schemas-
1) Staging schema - Data from all the operational data sources here raw clinical data . All the data in the staging schema has been dumped from this source to the working schema.It generally involves transferring the data from the source files to the respective tables in the staging database.
2) Working schema-From staging to working schema we will clean up the data
Data is not lost by cleaning or filtering but it is made available in some other usable form.We can't use the data as such in raw form. It involves transferring the data.

*Transformations performed in kettle-*

Different mappings are done in kettle .First the data from the csv files are transferred to the staging database and then data is processed from tables in staging schema to the tables in the functional schema in order to create a data warehouse.


**Figure5: Pentaho Data Integration**

Usually a significant amount of transformation of data occurs at the passage from the operational level to the data warehouse level.The transformations will read records from a input .CSV files, and then it will filter them out and write output to a separate table. The records which will pass the validation rule will be spooled into a text file and the ones that won't will be redirected to the rejects link which will place them in a different text file. Assuming that the Spoon

application is installed correctly, the first thing to do after running it is to configure a repository. Once the ‘Select a repository’ window appears, it’s necessary to create or choose one. A repository is a place where all Kettle objects will be stored here MySQl database has been installed .To create new repositories click the ‘New’ button and type in connection parameters in the ‘Connection information’ window.There are some very useful options on the screen, one is ‘Test’ which allows users to test new connections and the other is ‘Explore’ which lets users browse a database schema and explore the database objects.

After clicking the ‘Create or Upgrade’ a new repository is created. By default, an user with administrator rights is created – it’s login name is *admin* and the password is also *admin*. If a connection with repository is established successfully, a Spoon main application window will show up.

To design a new transformation which will perform the tasks described above it’s necessary to take the following steps:

1) Click the ‘New transformation’ icon and enter its name in my case in the functional stage Final_Patient ,Final_Test are one of them and in case of staging Test, Patient are the names of transformations done.

2) Define a database connection. It is located in the left hand-side menu in the ‘Main tree’ area in the Database connections field func was one of the Connections in my case.

3) Drag and drop the following elements from the ‘Core Objects’ menu to the transformation design area in the center of the screen: Table Input (menu Output), and one Field Output table objects (menu Output).

A mapping is the Kettle solution for transformation re-use. For example if you have a complex calculation that you want to re-use everywhere, you can use a mapping.A mapping is also called a sub-transformation because it is a transformation just like any other with a couple of key differences: Every mapping needs a Mapping Input step to define the fields that are required for the mapping to work correctly.Every mapping needs a Mapping Output step to define the fields that are generated by the mapping.

Because of the static nature of a mapping, Previewing mapping makes no sense.

4) Edit the Table Input – choose a source database and define an SQL query which will return records to the transform flow. The ‘Preview’ option is usually very useful here as it shows the preview of the records returned from the database.

5) Next thing to do is to link the objects together. The links between elements are called Hops and they indicate which direction the transform flows go. Hops elements can be found, created and edited in the Main Tree section.The easiest way to create a Hop is to drag and drop a link between two objects with left SHIFT pressed.

6) The last thing to do is to change the text files output configuration(MySQL table here). Enter the names of the files and its extension in the properties window and if needed, adjust other text files specific options here.

7) Save and run the transform (menu -> Transformation -> Run or just press the F9 key). Please find below execution log entries for a correctly configured and run Spoon transform.

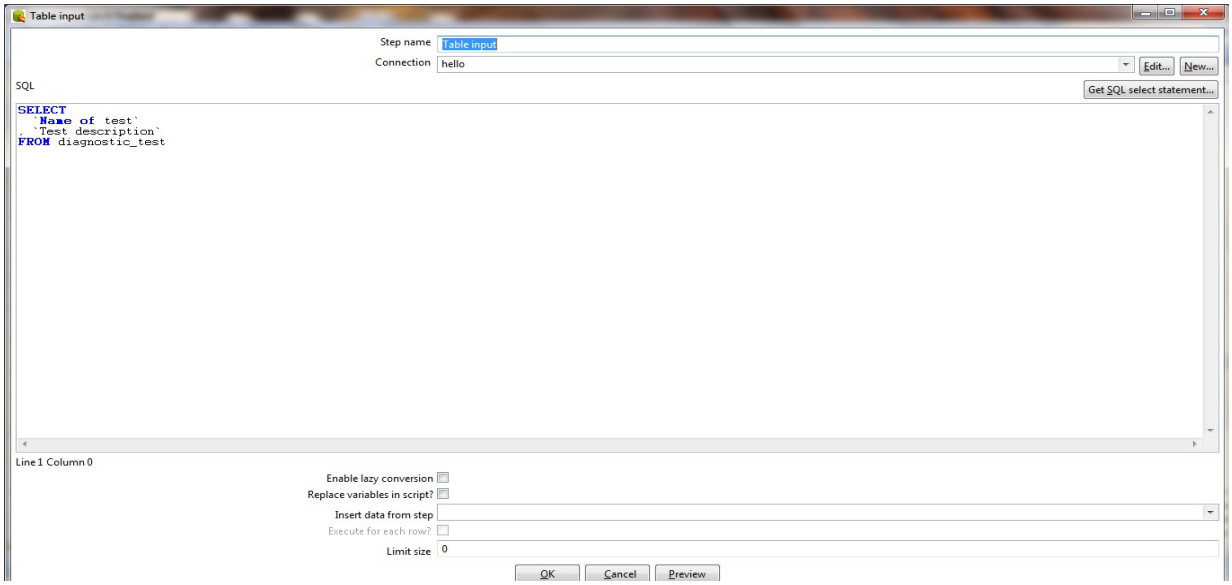



**Figure6: Using Mysql queries in Kettle to process the data.**

**Mappings in staging schema-**

**1)** Date-

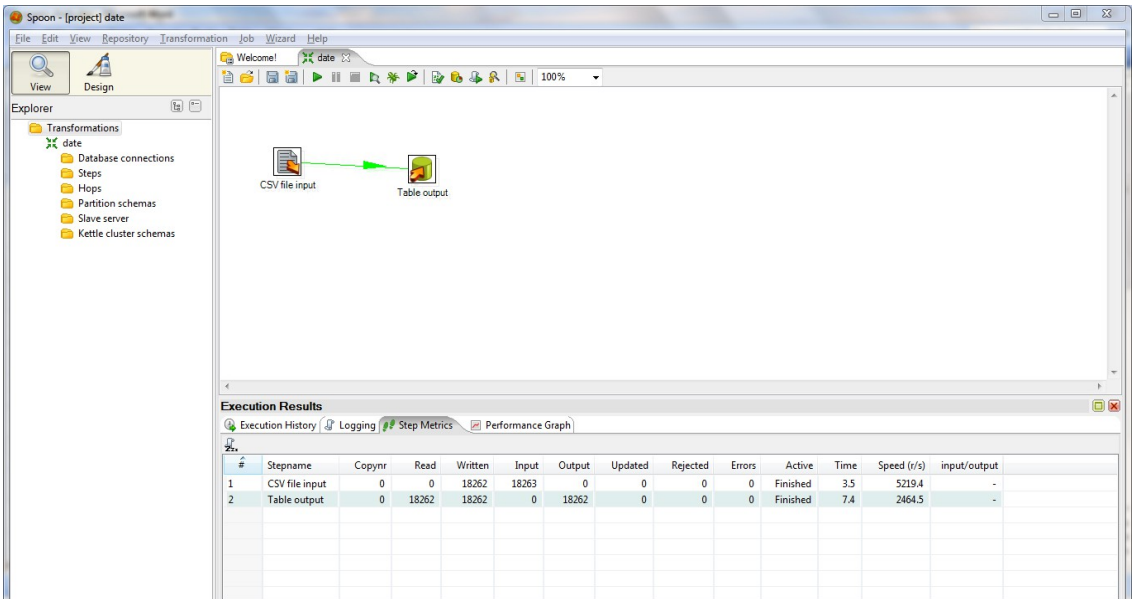



**Figure 7: Kettle transformation: Integrating the date data from csv file to the date table present in staging database.**

2)Hospital-

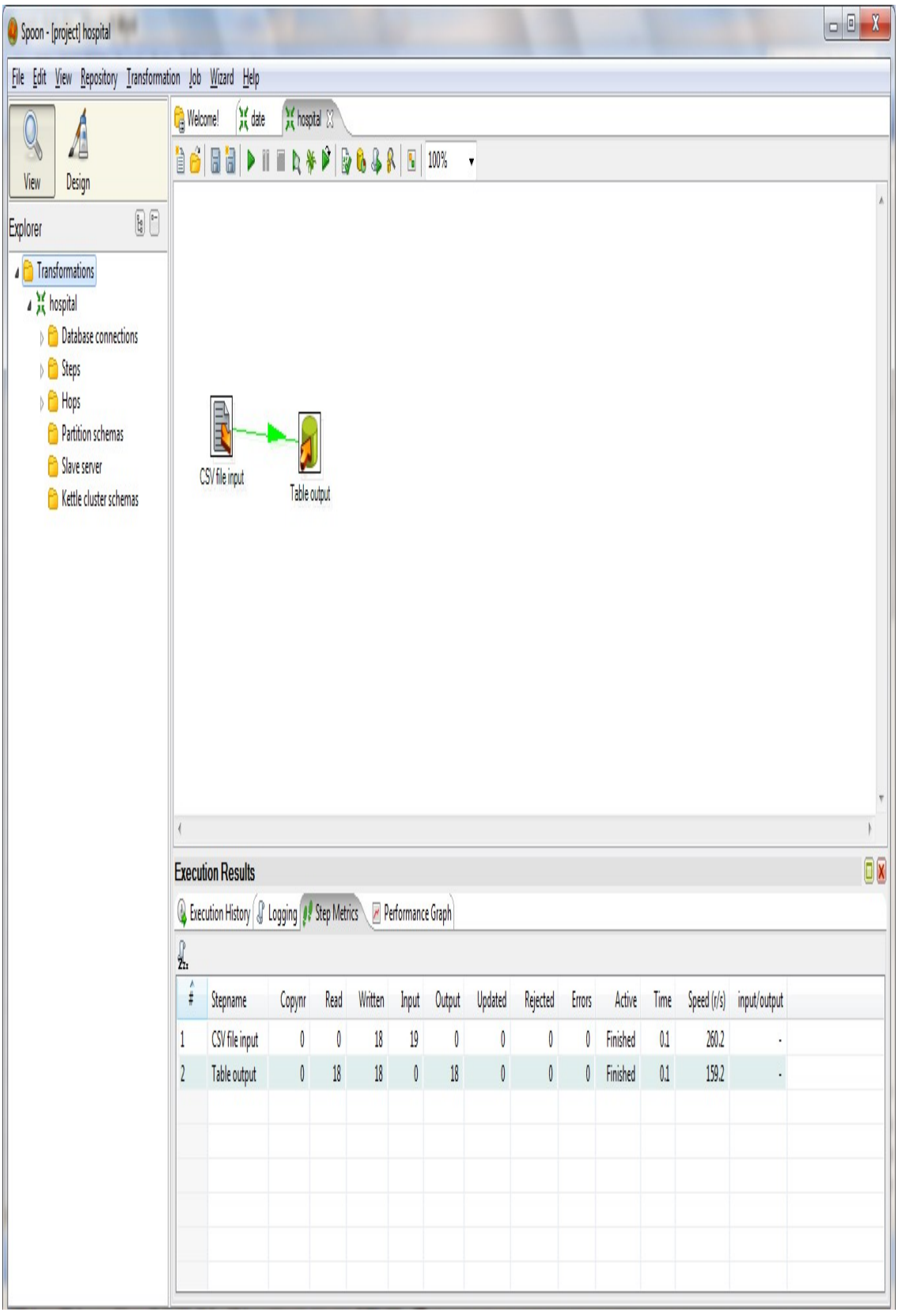



**Figure8: Kettle transformation:Integrating the data from csv file of hospital to the hospital table present in staging database.**

3)Disease-

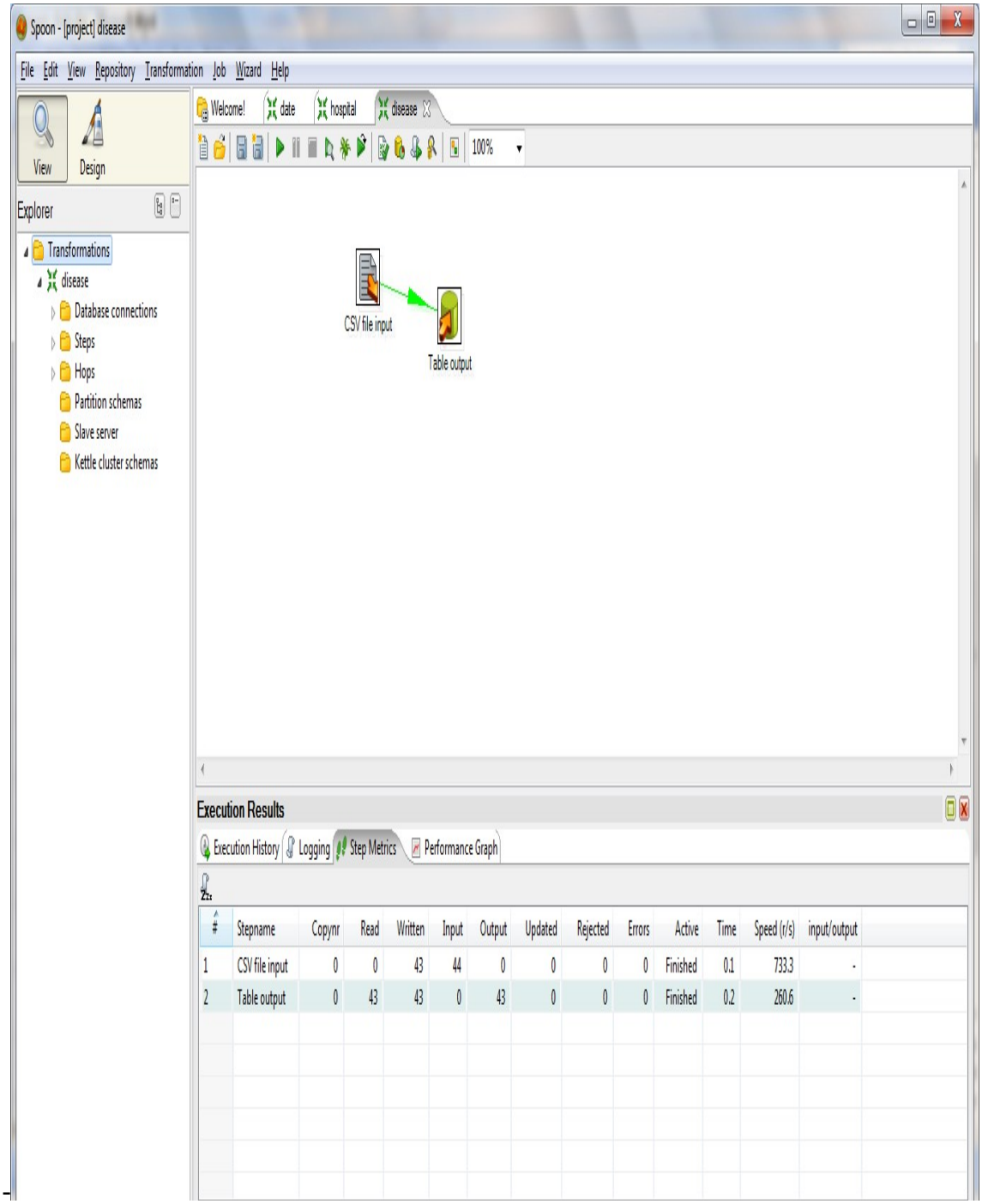



**Figure9: Kettle transformation:Integrating the data from csv file of disease to the table in staging database.**

4)Test-

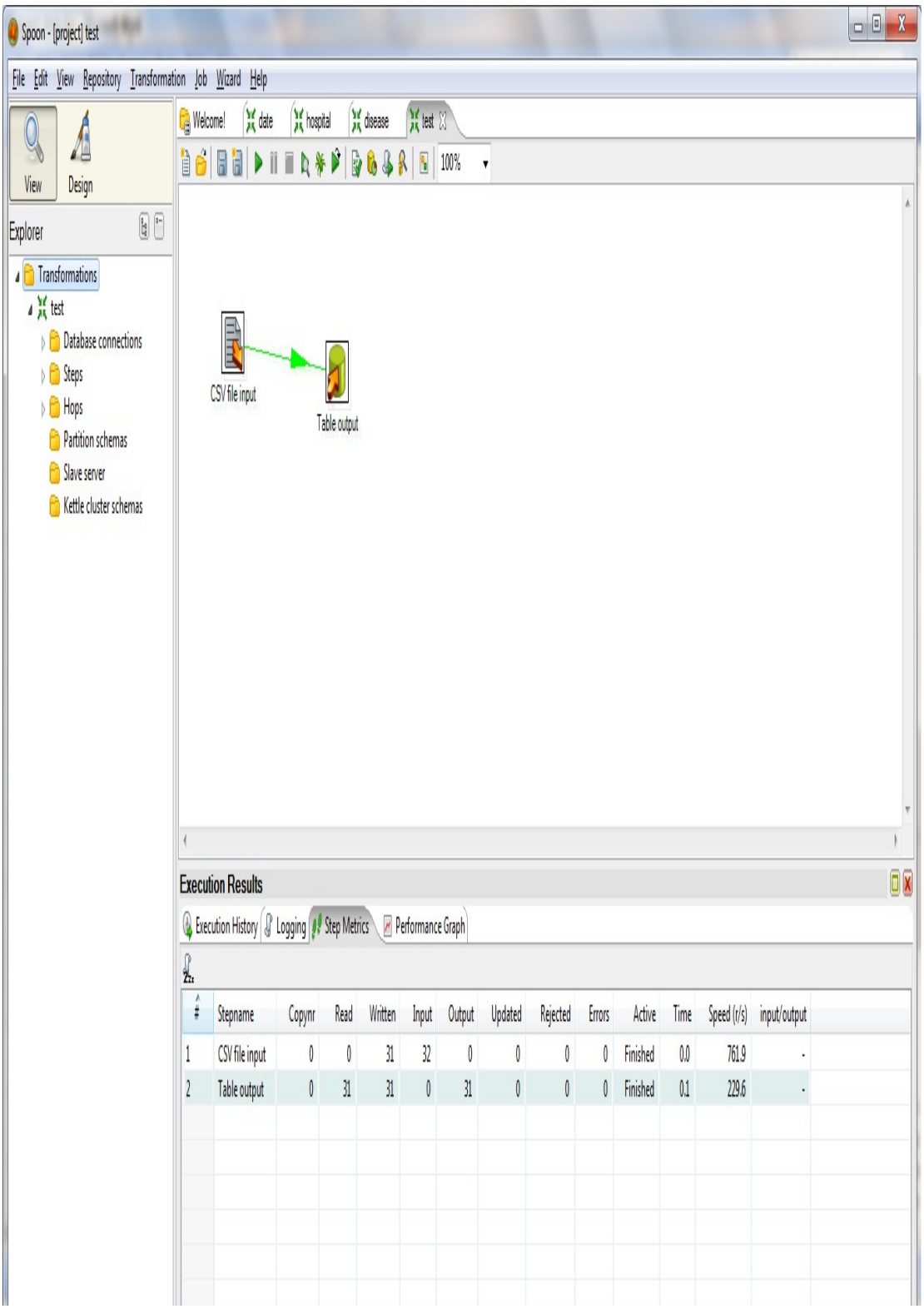



**Figure10: Kettle transformation:Integrating the data from csv file of test to the respective table in staging database.**

5) Patient-

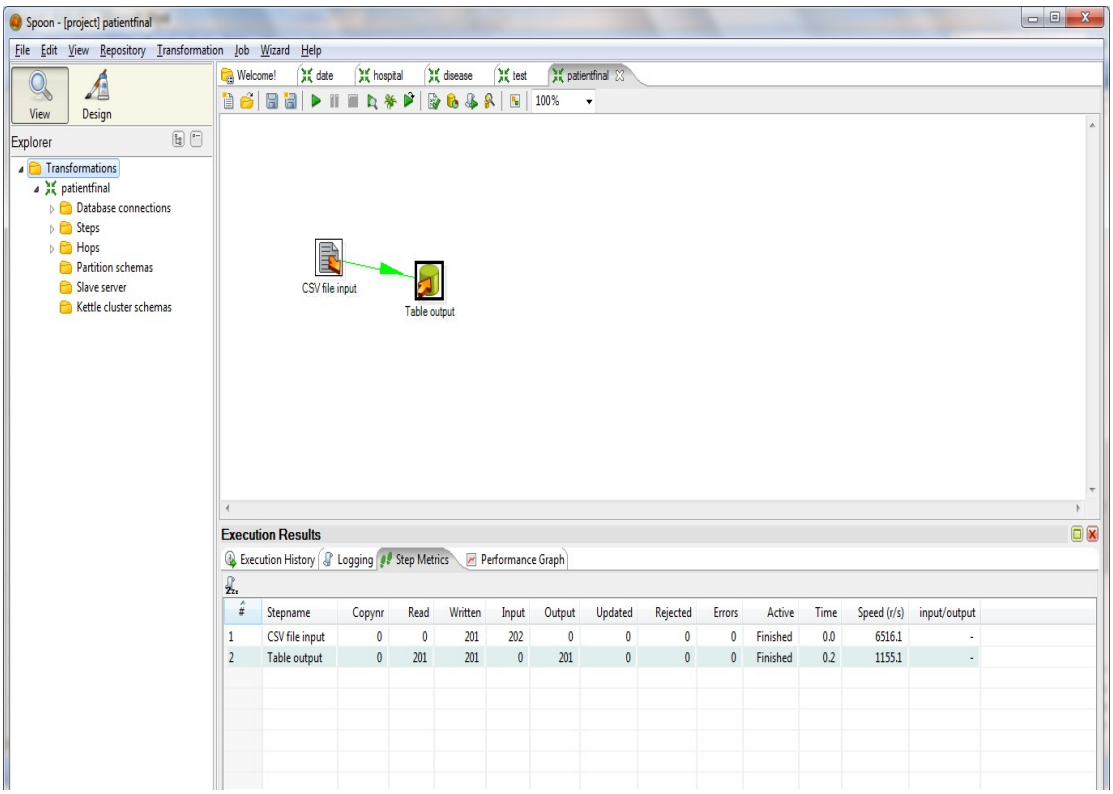



**Figure11: Kettle transformation:Integrating the data from csv file of patient to the respective table in staging database.**

**Mappings done in functional schema-**

1)Date-

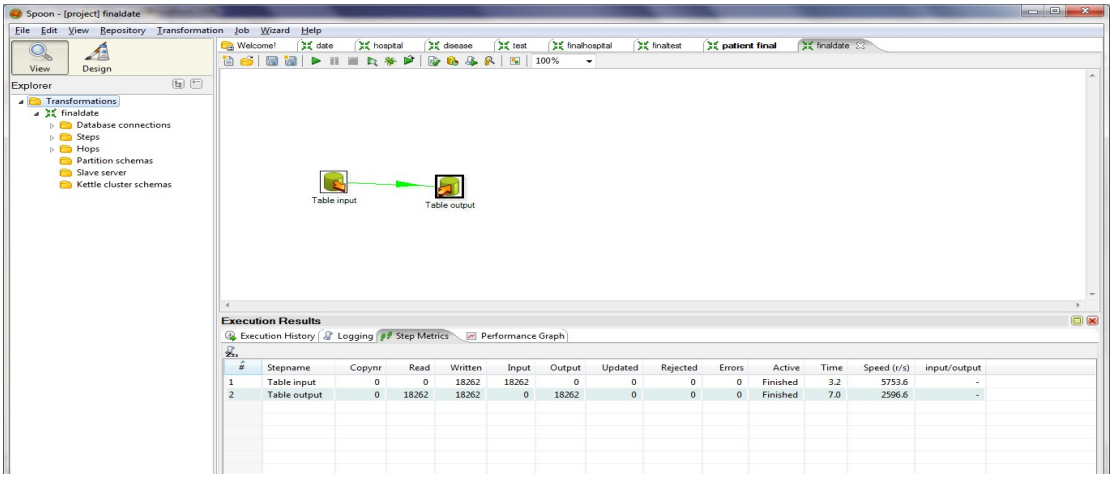



**Figure12: Kettle transformation:Integrating the data from the table in staging database(date) of the date to the dimensional table of date(dim_date) in functional database.**

2)Hospital-

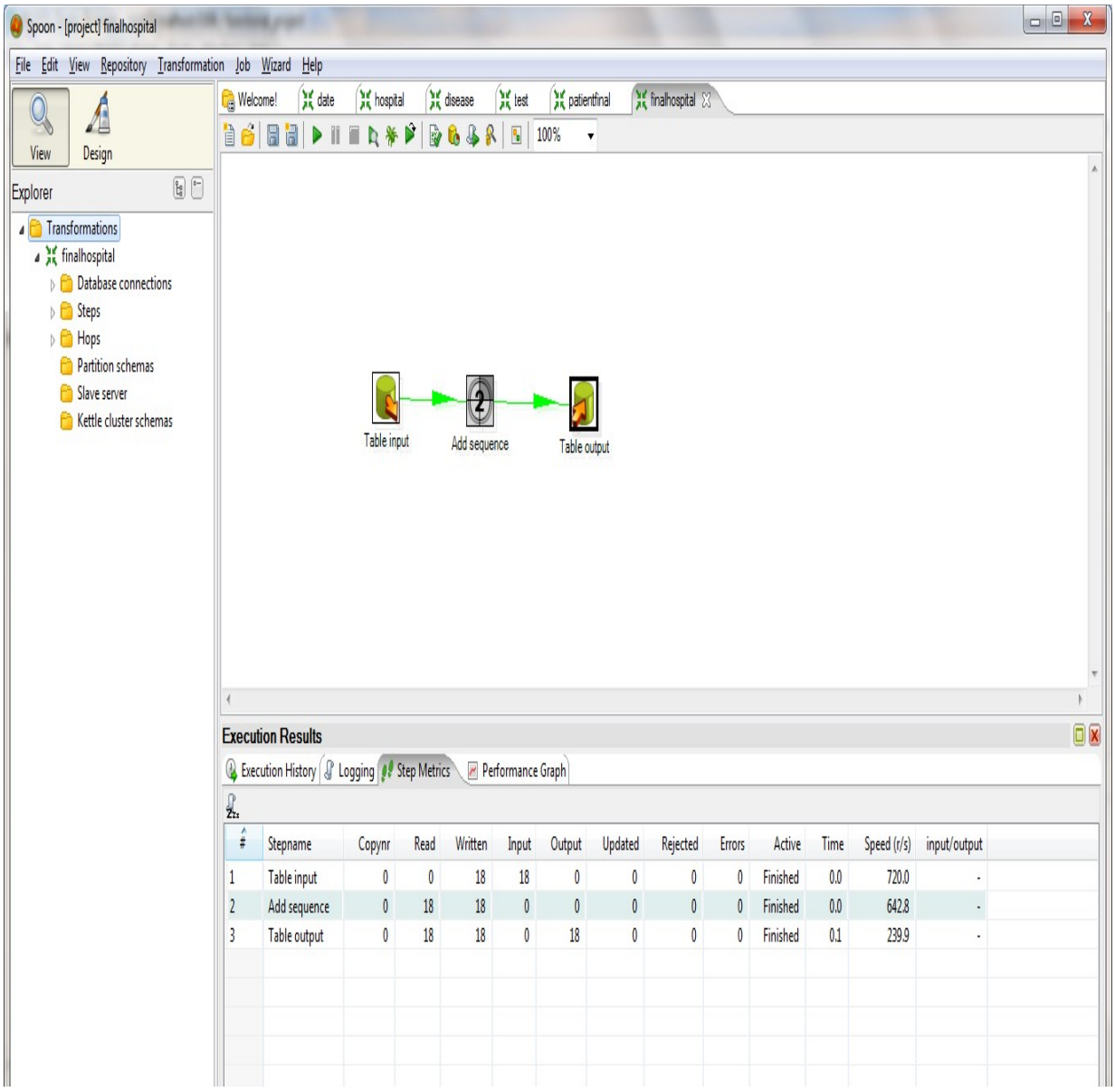

Spoon - [project] finalhospital
File Edit View Repository Transformation Job Wizard Help
View
Design
Explorer
Transformations
finalhospital
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Welcome!
date
hospital
disease
test
patientfinal
finalhospital
100%
Table input
Add sequence
Table output
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
# Stepname Copynr Read Written Input Output Updated Rejected Errors Active Time Speed (r/s) input/output
1 Table input 0 0 18 18 0 0 0 0 Finished 0.0 720.0 -
2 Add sequence 0 18 18 0 0 0 0 0 Finished 0.0 642.8 -
3 Table output 0 18 18 0 18 0 0 0 Finished 0.1 239.9 -

**Figure13:Kettle transformation:Integrating the data from the table in staging database(hospital) of the date to the dimensional table of date(dim_hospital) in functional database.**

**Note:In this case add sequence transformation is used in order to create or generate Hospital_ID in the data.**

3) Disease-

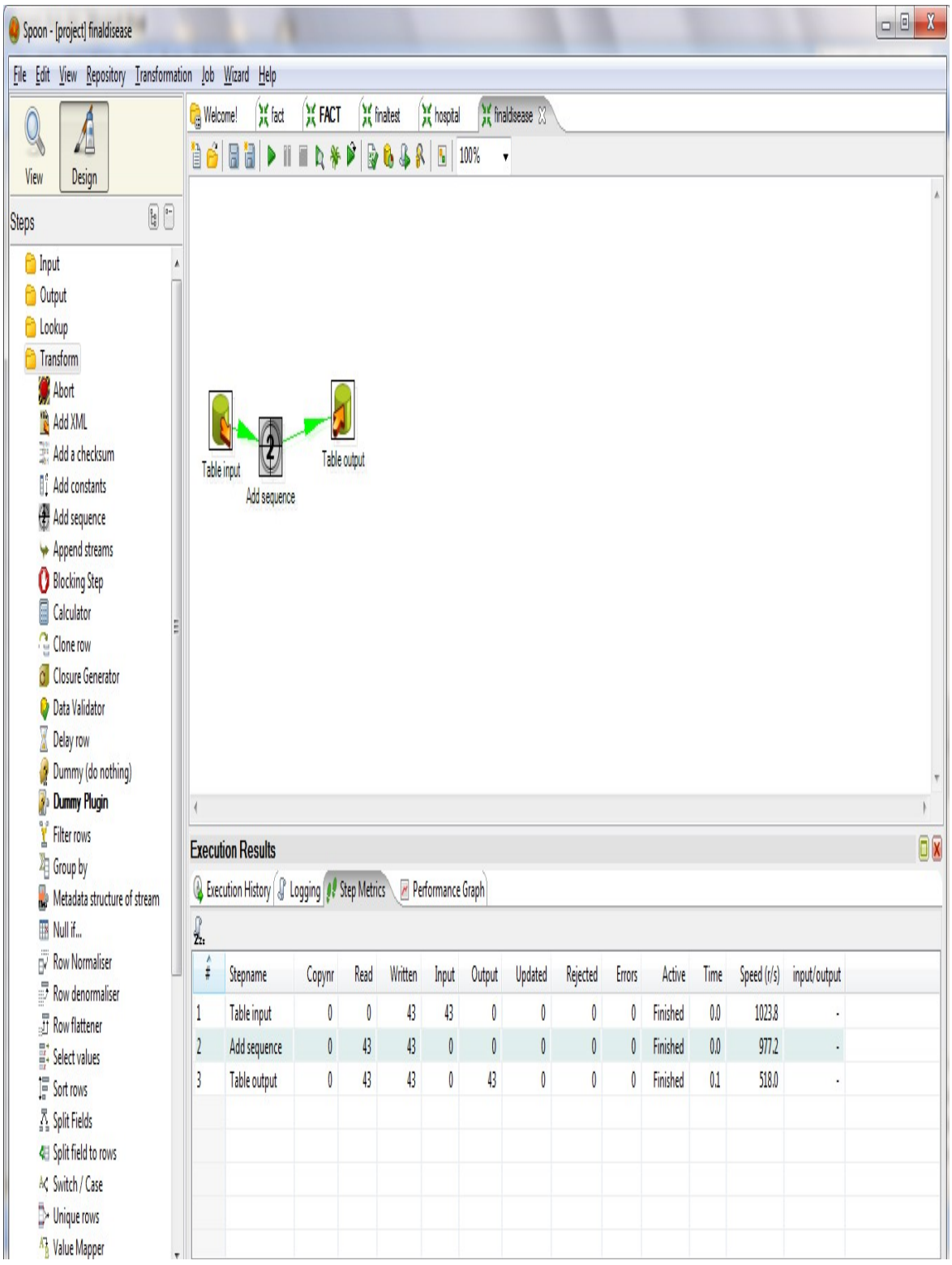



**Figure14: Kettle transformation:Integrating the data from the table in staging database(disease) of the date to the dimensional table of date(dim_disease) in functional database. Note:In this case add sequence transformation is used in order to create or generate Disease_ID in the data.**

4)Test

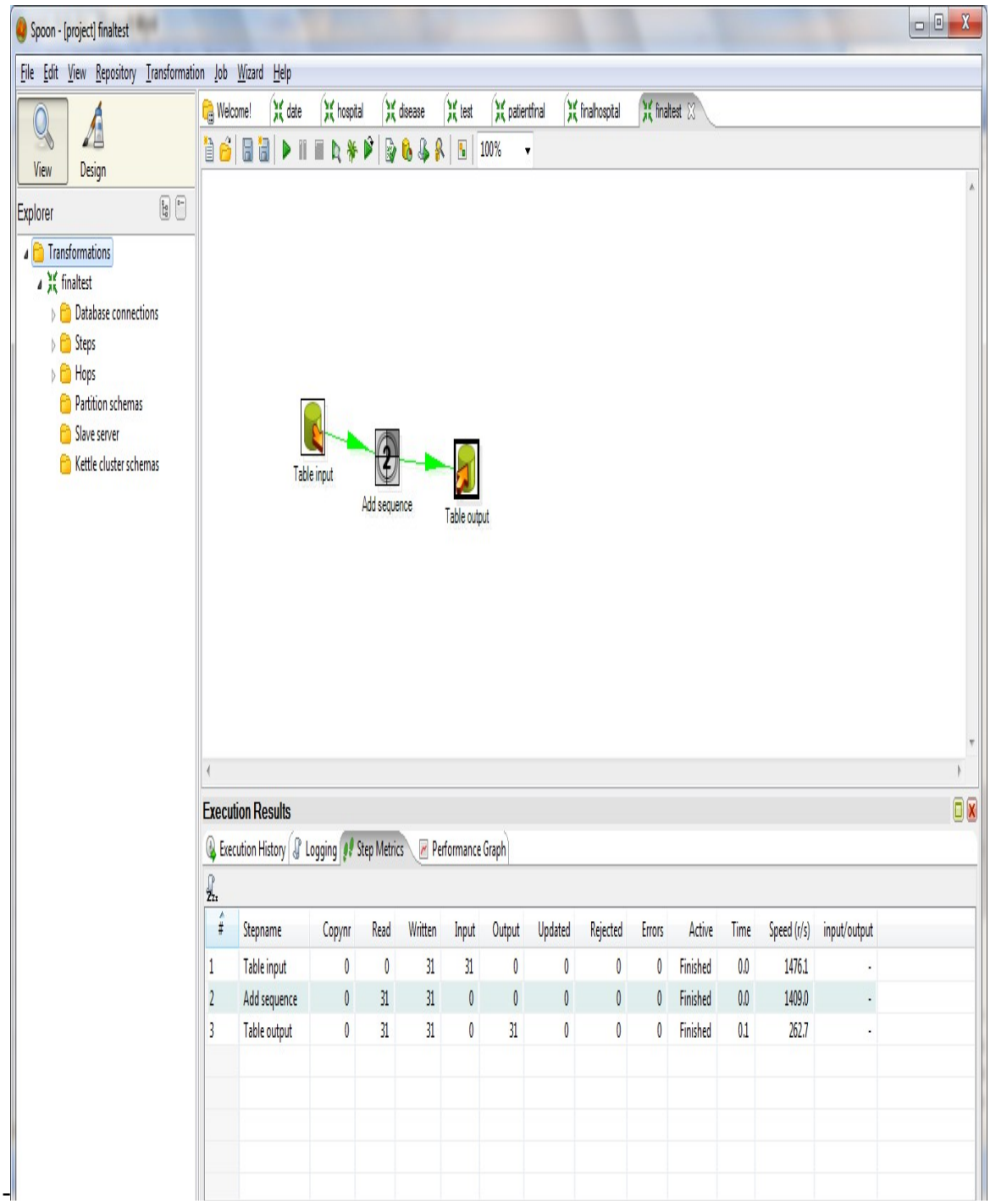



**Figure15: Kettle transformation:Integrating the data from the table in staging database(diagnostic_test) of the date to the dimensional table of date(dim_diagnostic_test) in functional database. Note:In this case add sequence transformation is used in order to create or generate Test_ID in the data.**

5) Patient-

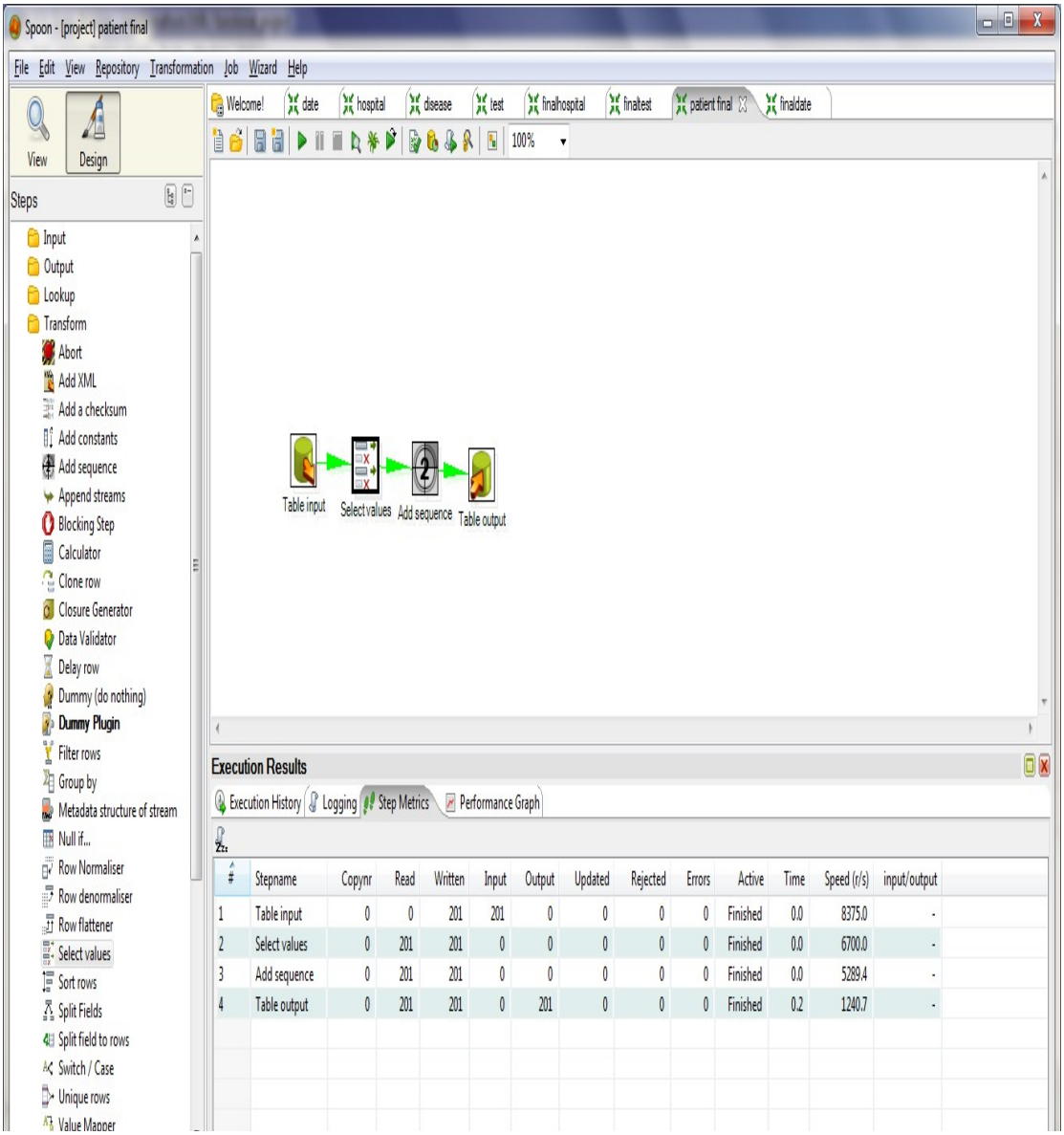



**Figure16: Kettle transformation:Integrating the data from the table in staging database(patient) of the date to the dimensional table of date(dim_patient) in functional database.**

**Note:In this case add sequence transformation is used in order to create or generate Patient_ID in the data and the select values to select particular columns from the table.**

6)Fact table-

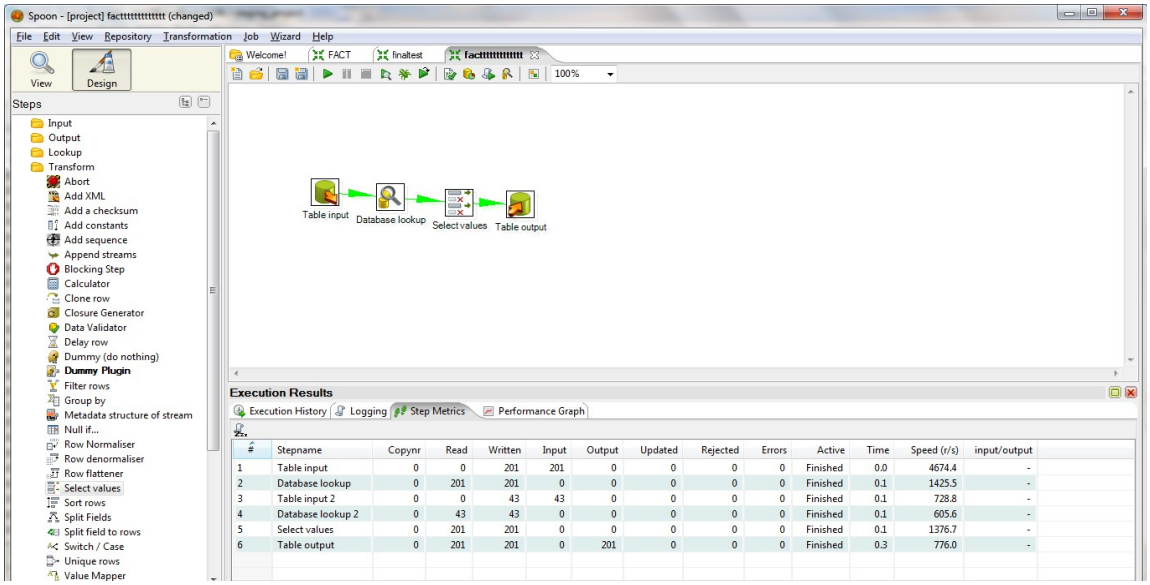



**Figure17: Loading the data into the fact table**

**Working schema-**

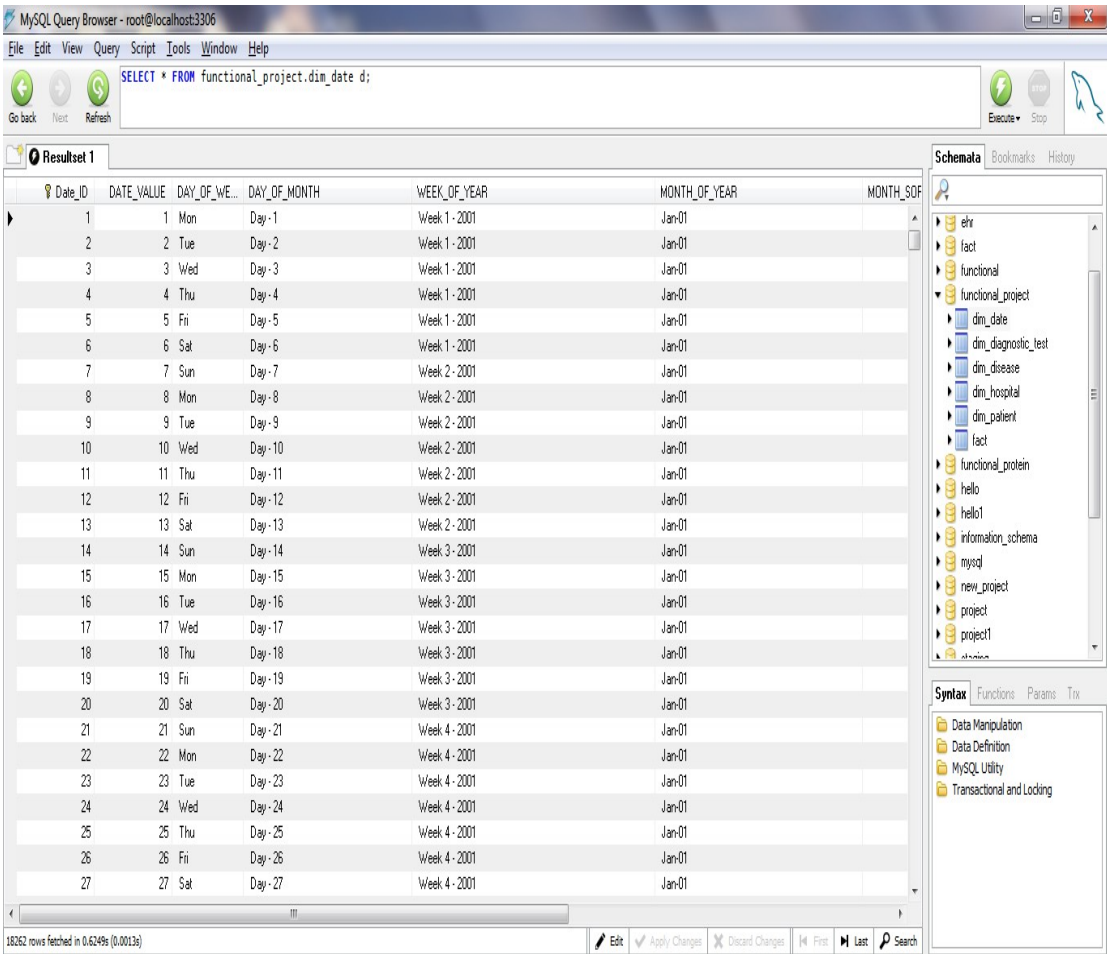



**Figure18: Dimension_date**

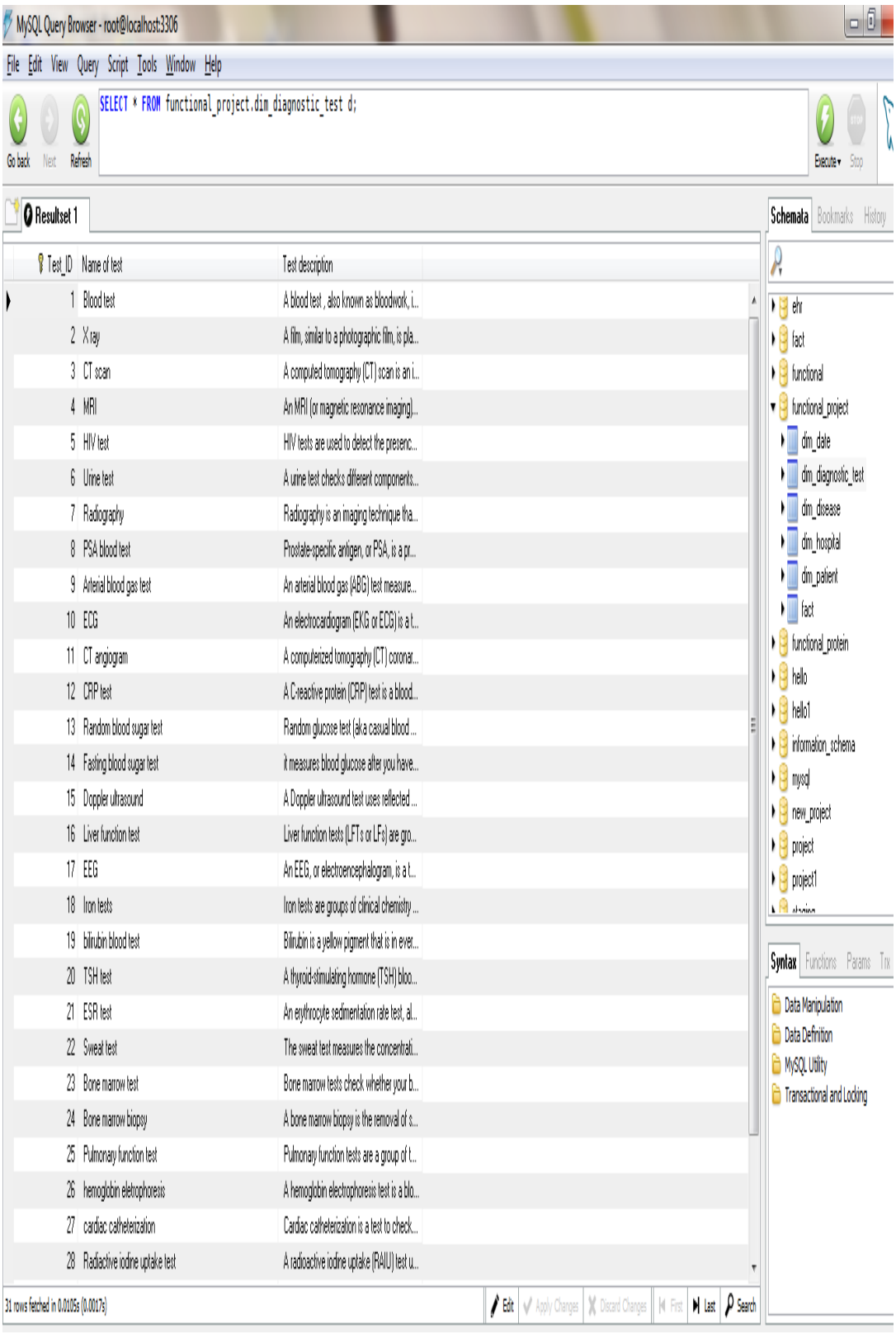



**Figure19: Dimension_Diagnostic_test**

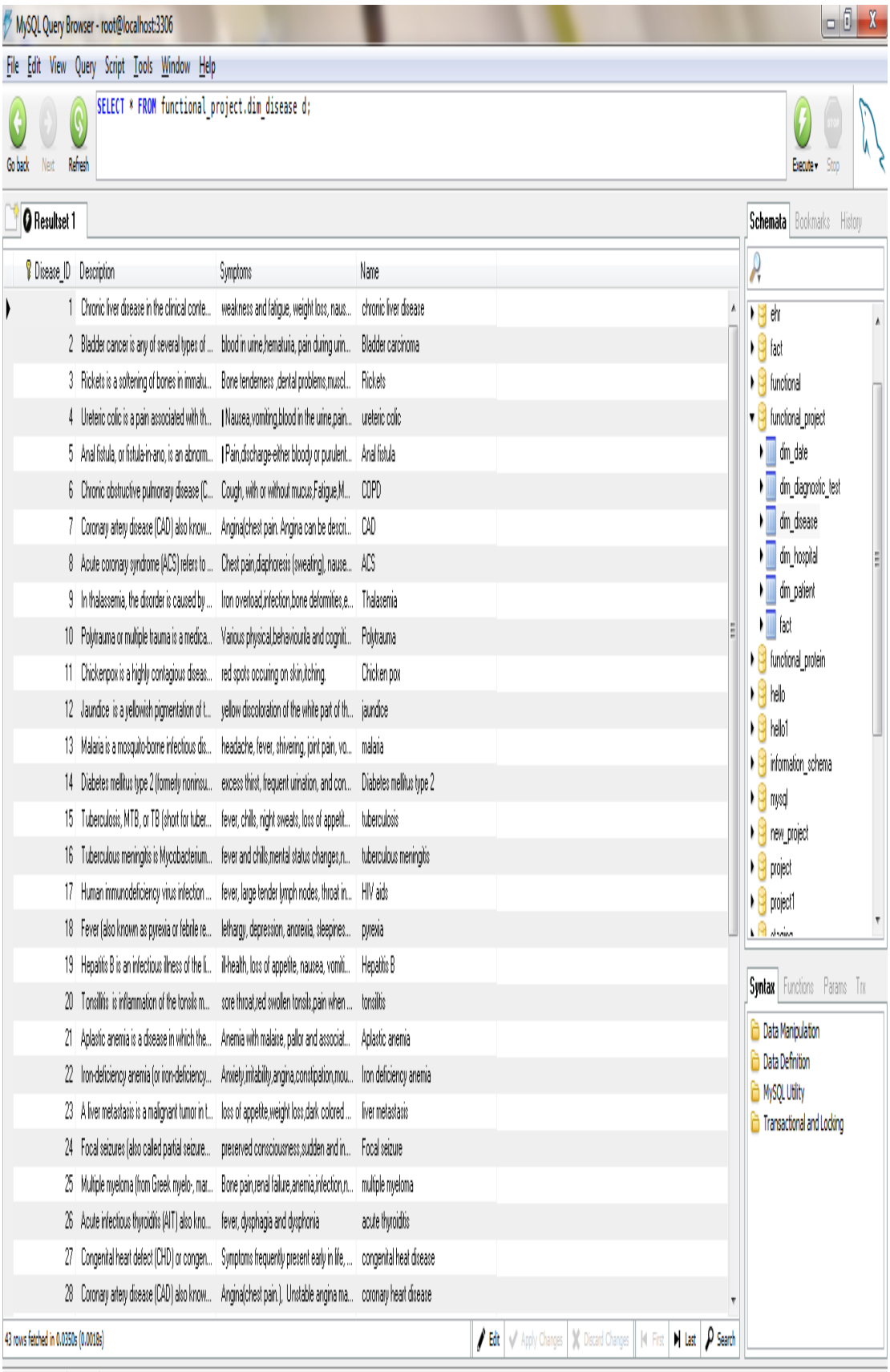



**Figure20: Dimension_disease**

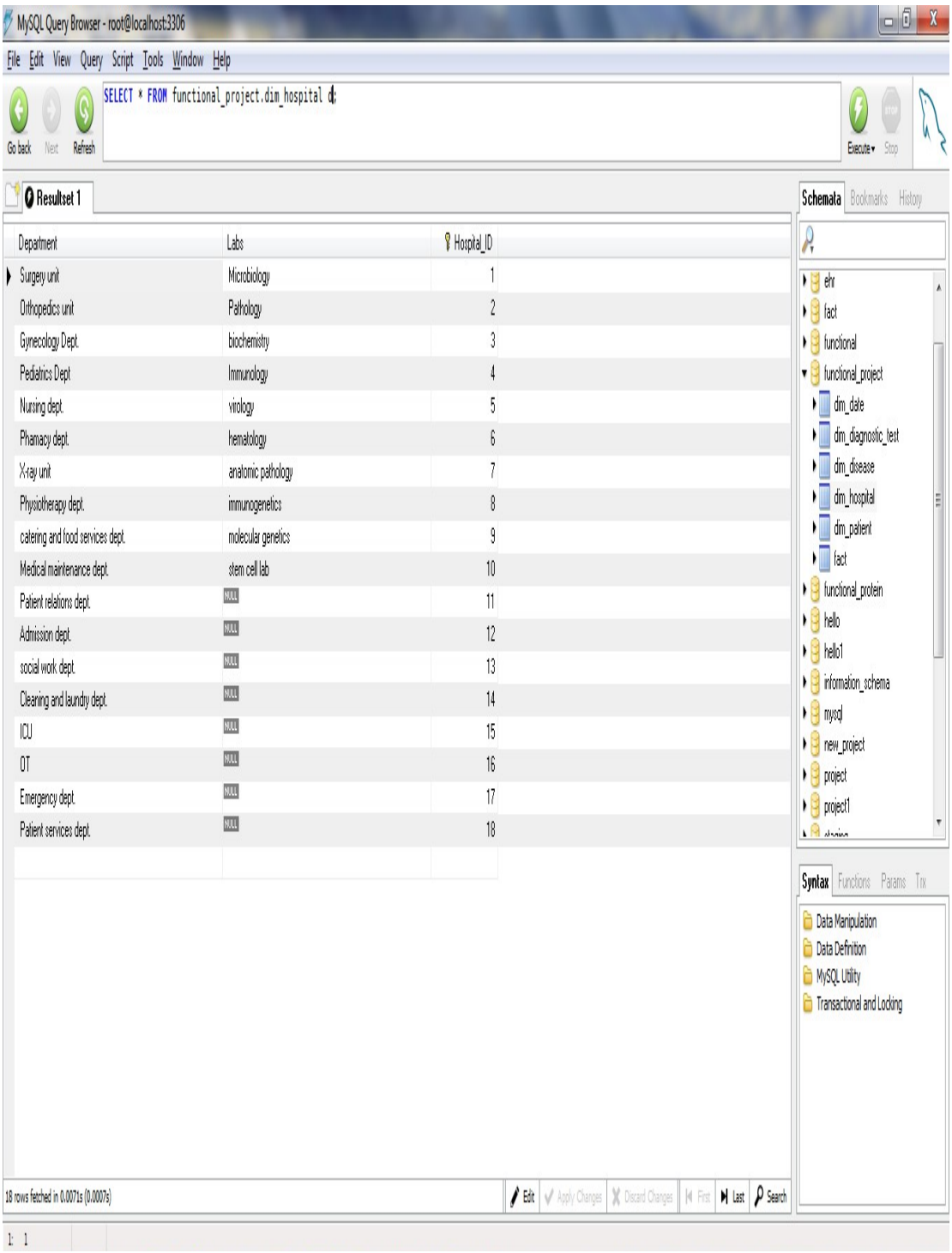



**Figure21: Dimension_hospital**

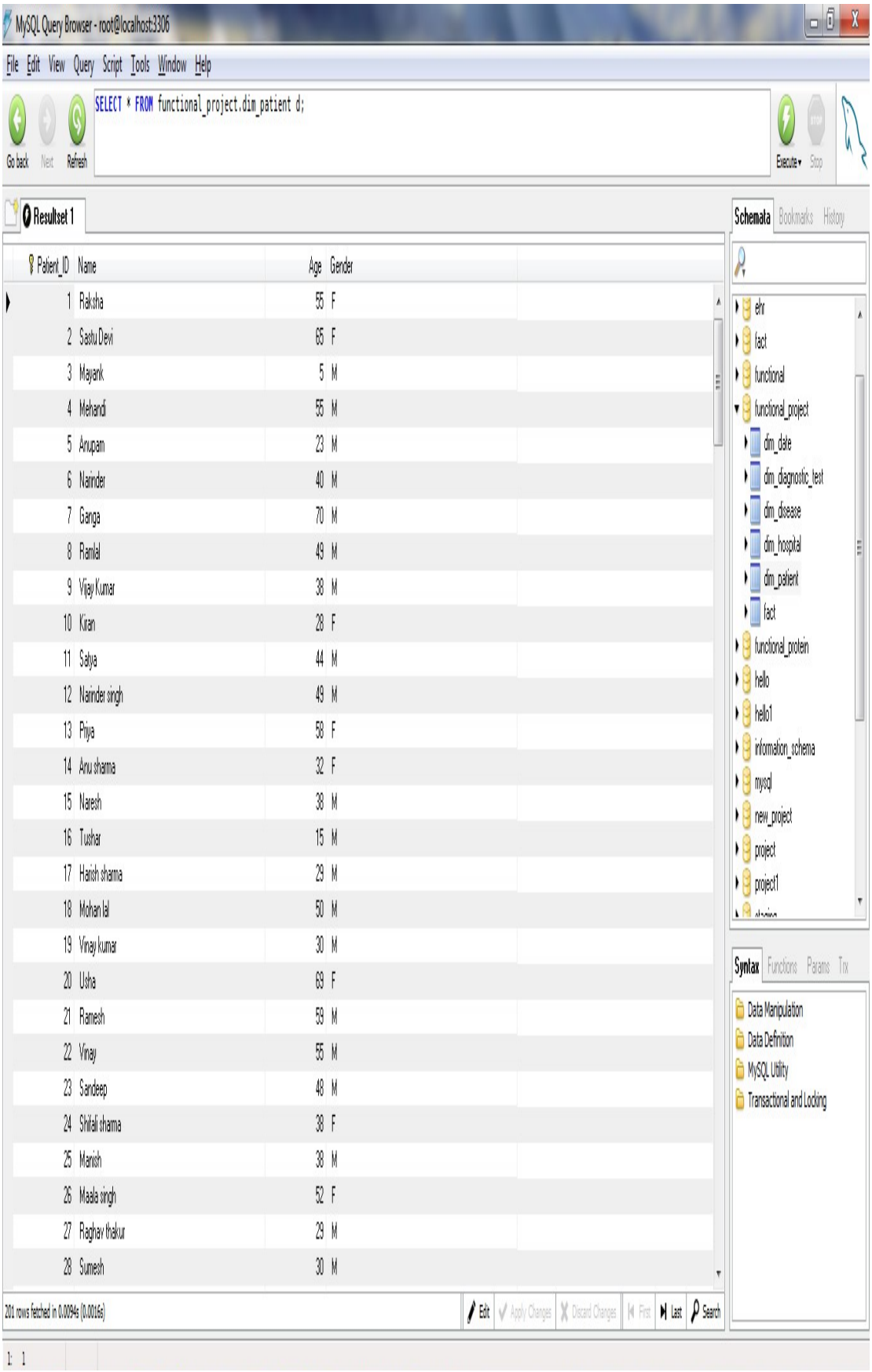



**Figure22: Dimension_patient**

# 4.CONCLUSION AND FUTURE WORK

A data warehouse has been developed for the patient care model by first developing the dimensional model which consists of facts which store the measureable quantities and the dimensions which store the descriptive attributes. Then data ware house is developed by making use of Mysql and kettle.Data warehouse helps in storing all the information related to the patient care model and helps in the management reporting process.

However in future data mining techniques can be applied to find the hidden patterns in the large clinical set which help in decision making process. On finding of these patterns many problems of the clinical data can be solved. New knowledge will be discovered which helps in making decision making process.

# APPENDIX-

Syntax of Mysql queries used-

Creation of database in staging schema named as staging_project

Create database staging_project;

Use staging_project;( which will highlight this database and all the tables are created under this database)

1) Creation of table patient in staging database-

Create table patient(Name varchar(50),Age int,Gender varchar(50));

2) Creation of table disease –

Create table disease(Name varchar(70),Description(70),Symptoms(50));

3) Creation of table hospital-

Create table hospital(Department varchar(50),labs varchar(30));

4)Creation of table diagnostic_test

Create table diagnostic_test( Name_of _test varchar(50),Test _description varchar(70));

5)Creation of table date

Create table date(Date_ID int,Date_value date,Day_Of _Week date,Day_of_Month date,Week_Of_Year date,Month_Of_Year date,Month_Sort_Value date,Week_Sort_Value date,Qtr_Of_Year date,Qtr_Sort_Value date,Calender_Year date));

Then we use functional database named as functional_project;

Use functional_project;

1)Creation of dimensional table dim_patient-

Create table dim_patient(Patient_ID int primary key,Name varchar(50),Age int,Gender varchar(50));

2)Creation of dimensional table dim_hospital-

Create table dim_hospital(Hospital_ID int primary key,Department varchar(50),labs varchar(30));

3)Creation of dimensional table dim_diagnostic test(Test_ID int primary key, Name_of _test varchar(50),Test _description varchar(70));

4)Creation of dimensional table dim_disease-

Create table dim_disease (Disease_ID int primary key,Name varchar(70),Description(70),Symptoms(50));

Creation of fact table-

```
Create table fact(Patient_ID int,foreign key fk1(Patient_ID)references
dim_patient(Patient_ID),Disease_ID int,foreign key fk2 (Disease_ID)references
dim_disease(Disease_ID),Test_ID int,foreign key fk3(Test_ID)references
dim_diagnostic_test(Test_ID),Hospital_ID int,foreign key fk4(Hospital_ID)references
dim_hospital(Hospital_ID),
Date_ID int ,Date_ID_Discharge int,foreign key fk5(Date_ID)references dim_date(Date_ID));
```

## REFERENCES-


[1]Giuse G. *KR: Constraint-Based Knowledge Representation.* Technical Report CMU-CS-89-142, Carnegie Mellon University, April 1989.

[2] Bailey J. A Knowledge-based System to provide advice and explanation in post-operative care. in *Trends in Computer Assisted Education,* ed. Lewis and Tagg, London, Blackwell Scientific Publications, 1989.

[3] Bailey J. *Explanation-Giving Systems for Learning Decision-Making Skills*. Ph.D. Thesis 1989, Computer Based Learning Unit, University of Leeds.

[4]Kelleher G and Bailey J. An Explanation Driven Architecture for a Knowledge-based System in Postoperative Care. in *Proceedings of AIME 89* Medical Informatics Vol. 38, 1989, Springer Verlag.

[5]Myers B A. *The Garnet Toolkit Reference Manuals*, CMU-CS-89-196, Carnegie Mellon University, March 1990.

[6]Sawar M J, Brennan T G, Cole A J and Stewart J. POEMS (PostOperative Expert Medical System), in *Proceedings of IJCAI-91 one Day Workshop : "Representing Knowledge in Medical Decision Support Systems"*, Sydney, Australia, Aug. 1991.

[7]Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004;25:690-5.

[8] Zhu X. Semi-Supervised Learning Literature Survey. University of Wisconsin-Madison; 2007.

[9]Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. AAAI Press; 1996. p. 82-8.

[10]Krivda CD., Data-Mining Dynamite. Byte,(1995) Oct 95:97-102.

[11] Hedberg, SR., The Data Gold Rush. Byte, 1995;Oct :83-88.

[12] Burkett, ME., The Tertiary Center and Health Departments in Cooperation: The Duke University Experience. J Perinat Neonat Nursing,(1989) 2:11-19.

[13] Prather JC, Lobach DF, Hales JW, Hage ML, Fehrs SJ, Hammond WE. (1995) Converting a Legacy System Database into Relational Format to Enhance Query Efficiency. Proceedings Annual Symposium Computer Applications Medical Care,19:372-376.

[14] Greenfield L. ,The Data Warehousing Information Center. LGI Systems Incorporated.(1996)

[15]Woolery, L, and Grzymala-Busse, J. (1994) Machine learning and preterm birth risk assessment. Journal of the American Medical Informatics Association. 1(6):439-446.

[16]D. McMichael, "Data fusion for vehicle-borne mine detection," in *Proc. 1st IEE Conf. Detection of Abandoned Land Mines*, Edinbourgh, UK, 1996, pp. 167–171.

[17] H. Pan and McMichael, “Information fusion, causal probabilistic network and Probanet,” in *Proc. 1st Int. Workshop Image Analysis and Information Fusion*, Adelaide, Australia, 1997, pp. 445–458.
[18] A. Assi, “Data fusion for medical applications,” in *Proc. Fusion’98*, Las Vegas, NV, 1998, pp. 447–450.
[19] T. W. Liao, Z. Zhang, and C. R. Mount, “Similarity measures for retrieval in case-based reasoning systems,” *Appl. Artif. Intell.*, vol. 12, pp. 267–288, 1998.
[20] *Coronary Heart Disease: An Epidemiological Overview*. London,U.K.: HMSO, 1994.
[21] Cohen, D. Hudson, and P. Deedwania, “Combination of chaotic and neural network modeling for diagnosis of heart failure,” in *Proc. Int. Conf. Computers and Their Applications*, 1997, pp. 254–257.
[22] Z. Shen, M. Clarke, R. Jones, and T. Alberti, “A neural network approach to the detection of coronary artery disease,” in *Proc. IEEE Computers in Cardiology*, vol. 20, pp. 221–224, 1993.
[23]P. Lopes, R. Mitchel, and J. White, “The relationships between respiratory sinus arrhythmia and coronary heart disease risk factors in middle aged males,” *Automedica*, vol. 16, pp. 71–76, 1994.
[24] F. Azuaje, W. Dubitzky, P. Lopes, N. Black, K. Adamson, X. Wu, and J. White, “Predicting coronary disease risk based on short-term RR intervals measurements: A neural network approach,” *Artif. Intell. Med.*, vol. 15, pp. 275–297, 1999.
[25]K. Anderson, P. Wilson, P. Odell, and W. Kannel, “An updated coronary risk profile: A statement for health professionals,” *Circulation*, vol. 83,pp. 356–361, 1991.
[26] B. Fritzke, “Growing cell structures-a self-organizing network for unsupervised and supervised learning,” *Neural Networks*, vol. 7, pp.1441–1460, 1994.
[27] B. Ripley, in *Pattern Recognition and Neural Networks*. Cambridge,U.K.: Cambridge Univ. Press, 1996, pp. 311–326.
[28] T. Kohonen, *Self-Organizing Maps*. Heidelberg, Germany: Springer-Verlag, 1995.