PROJECT 1

Data Warehouse / OLAP System

Prachi Gokhale(50096829)
Pradnya Kulkarni(50096109)
Mahesh Jaliminche(50097738)

Table of Contents

Page #

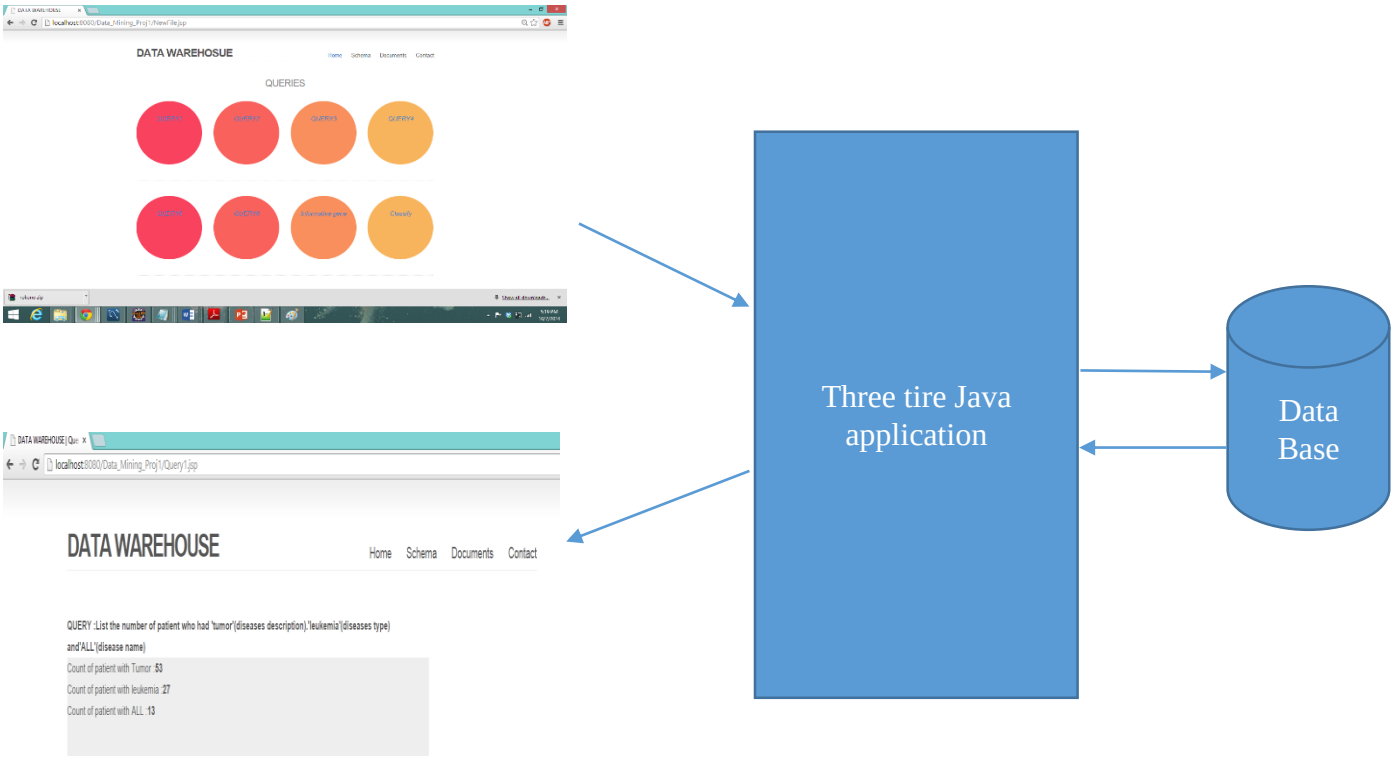
PROJECT DESCRIPTION:

In this project we need to implement a clinical and genomic data warehouse based on the schema design using any database system. A good data warehouse should satisfy the following requirements:

- Support regular and statistical OLAP operations
- Be robust to potential change in future.
- Support knowledge discovery

SYSTEM COMPONENT:

The system consist of two main components: A database which store data warehouse data and an java application which takes Input from user and display result



IMPLEMENTATION:

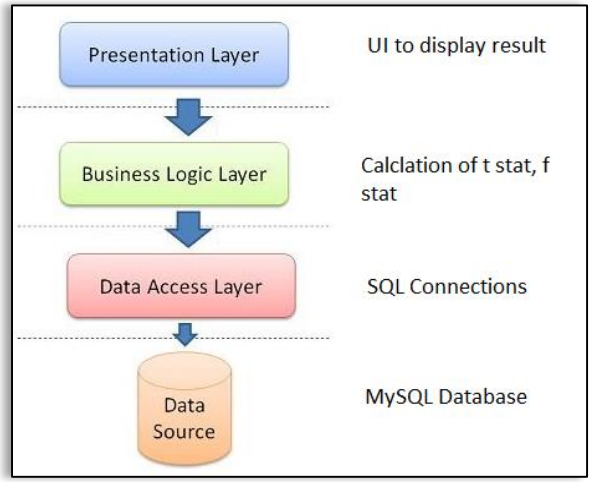


Fig: Data warehouse Application

Part 1: Implement data warehouse schema in a Database system

In this part we design our own schema for the data warehouse and load the input data into our warehouse. We have used **MySQL** database to implement our data warehouse.

Input Data: The original data is provided in a plain text files under the directory /projects/azhang/cse601/Data_For_Project1.

Data - Set:
There are five data spaces:

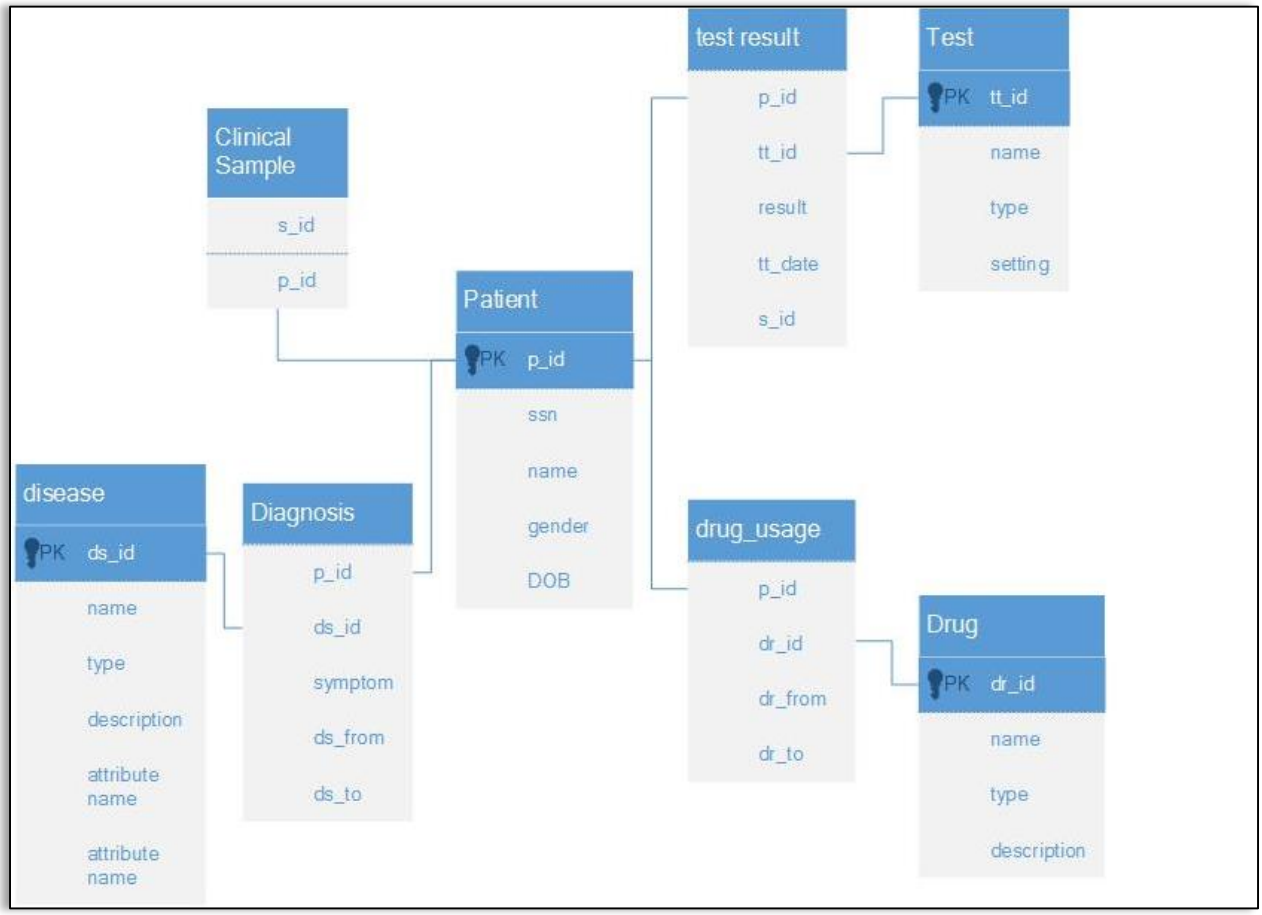
1. Clinical data space
2. Sample data space
3. Microarray and proteomic data space
4. Gene data space
5. Experiment data space

SCHEMA Implemented: (BioStar)

A BioStar fact schema: is a quadruple F= (C, D, M, S), where C is the central entity schema, D is a set of dimension schemas, M is a set of measure schemas and S is a set of summarisability constraints. The central entity is viewed as a special dimension, which is associated with every fact measure. For example, in the clinical data space, Patient is the central entity; and in the gene data space, Gene Sequence is the central entity.
A measure schema: Mj∈M, is a triple (Am, As, Dm), where Am is a set of attributes called measures, As is a set of supporting attributes for the measures (e.g., single- or bi-temporal support) and Dm is a set of dimensions that are associated with the measures. In a relational database, a

measure schema may be implemented as a separate table, which we call it an m-table. An m-table is associated with the central entity and one or more dimensions. Note that Am can be an empty set. In such cases, the m-table just keeps the relationships among the central entity and the associated dimensions, and the records in the m-table may be used for counting occurrences instead of numeric aggregation.

Schema Design for Clinical Data Space



Advantages:

1. The BioStar model has the property of great extensibility, which is important for some fast-evolving data spaces such as the clinical and gene data spaces. In these data spaces, existing dimensions may need to be modified, and new dimensions may be added over time. BioStar's extensibility is realized by storing different measures in separate m-tables

2. The many-to-many relationships between the central fact entity and dimensions are handled using the m-tables. For each of the many-to-many relationships, an m-table is created.

3. Uncertain relationships between the central entity and dimensions may be kept in the m-tables. An additional field may be included in the m-table to specify if a relationship instance is uncertain.

Schema Design for Complete Data Space

Part 2: Support Regular and statistical OLAP Operation

Operations on Data warehouse:

1. Queries

   In our biomedical data warehouse, the commonly used aggregation operators are SUM, AVG, MAX, MIN and COUNT.

2. Statistical Operations

   T Statistics, F Statistics, CORRELATION, T-TEST, ANOVA are the statistical operators that are widely used in biomedical research.
   The CORRELATION operator is used to compute the Pearson or Spearman correlation coefficient between two random variables.
   T-TEST is used to determine if there is a significant difference between two random variables by computing the t-statistic and ANOVA (analysis of variance) is used to test whether there are differences between any pairs of random variables.

Part 3: Support Knowledge Discovery

In this part we use our data warehouse and the OLAP operations to support Knowledge discovery.

1) Given a specific disease find the informative gene

Algorithm:

- Find all the patient with 'ALL' (group A), while the other patient serve as the control (group B).
- For each gene, calculate the t-statistics for the expression values between group A and group B
- If the p-value of the t-test is smaller than 0.01, this gene is regarded as **informative gene**

2) Given a new Patient Pn, we want to predict weather he/she has 'ALL'

Algorithm:

- Find the informative gene w.r.t. 'ALL'.
- Find all the Patient with 'ALL' (group A).
- For each patient Pa in group A, calculate the correlation Ra of the expression values of the informative gene between Pn and Pa.
- Patient without 'ALL' serve as control (group B)
- For each patient Pb in group B, calculate the correlation Rb of the expression values of the informative gene between Pn and Pb.
- Apply t-test on Ra and Rb, if p-value is smaller than 0.01, the patient is classified as 'ALL'.

RESULT:

1. Queries

1) Query 1: List the number of patients who had "tumor", "leukemia" and "ALL", respectively.

Disease Description	Number of Patients
tumor	53
leukemia	27
ALL	13

2) Query 2: List the types of drugs which have been applied to patients with "tumor".

Types of drugs applied to patients with "tumor" are 20

Drug Type 014	Drug Type 011
Drug Type 016	Drug Type 017
Drug Type 019	Drug Type 007
Drug Type 013	Drug Type 003
Drug Type 004	Drug Type 012
Drug Type 009	Drug Type 005
Drug Type 006	Drug Type 015
Drug Type 001	Drug Type 002
Drug Type 020	Drug Type 008
Drug Type 010	Drug Type 018

3) Query 3: For each sample of patients with "ALL", list the mRNA values (expression) of probes in cluster id "00002" for each experiment with measure unit id = "001".

Number of records (mRNA values expression) retrieved is 325

4) Query 4: For probes belonging to GO with id = "0012502", calculate the t statistics of the expression values between patients with "ALL" and patients without "ALL".

t-stat= 0.98

5) Query 5: For probes belonging to GO with id="0007154", calculate the F statistics of the expression values among patients with "ALL", "AML", "colon tumor" and "breast tumor".

F-stat =3.13

6) Query 6: For probes belonging to GO with id="0007154", calculate the average correlation of the expression values between two patients with "ALL", and calculate the average correlation of the expression values between one "ALL" patient and one "AML" patient.

0.1435443475016023
-0.003475600831930593

2. Knowledge Discovery

Use your data warehouse and the OLAP operations to support knowledge discovery. Given a specific disease, find the informative genes.

Disease Description	Number of Informative Genes
ALL	38
AML	16
Colon cancer	3
Breast cancer	6

Number of informative genes for "ALL"

UID	UID
1 -- 16073088	20 -- 87592194
2 -- 58672549	21 -- 92443312
3 -- 40567338	22 -- 88257558
4 -- 31997186	23 -- 41333415
5 -- 43866587	24 -- 15295292
6 -- 94113401	25 -- 74496827
7 -- 75492172	26 -- 85557586
8 -- 83398521	27 -- 52948490
9 -- 48199244	28 -- 60661836
10 -- 24984526	29 -- 65772884
11 -- 31308500	30 -- 47276861
12 -- 37998407	31 -- 21633757
13 -- 58792011	32 -- 69156037
14 -- 28863379	33 -- 4826120
15 -- 88596261	34 -- 97606543
16 -- 45926811	35 -- 75434512
17 -- 11333636	36 -- 1433276
18 -- 41464216	37 -- 18493181
19 -- 13947282	38 -- 53478188

3. Classification

   Use informative genes to classify a new patient.

Test1 is classified as having ALL
Test2 is classified as having ALL
Test3 is classified as not having ALL
Test4 is classified as having ALL
Test5 is classified as having ALL

REFRENCES:

http://dev.mysql.com/downloads/installer/
http://dev.mysql.com/downloads/connector/j/5.0.html
http://commons.apache.org/proper/commons-math/apidocs/org/apache/commons/math3/
www.w3schools.com