Project 1:Data Warehouse/OLAP System

Nischala Manjunath – 50135265

Sai Srinath Sundar - 50134232

Data Warehouse Design and Implementation

The given input data was in the form of .txt files. This data was then cleaned appropriately. How the files were cleaned was dependent on each file itself. After cleaning the files were stored in .csv format to facilitate imports onto the Oracle Database.

The Schema implemented can be seen as a set of star schemas with the center the other stars being the fact tables. There are five such star schemas implemented with the corresponding five fact tables being clinical_fact, sample_fact , microarray_fact, gene_fact and experiment_fact.

Clinical_fact is linked to its corresponding dimension tables: drug, disease, patient, test.

Sample_fact is linked to its corresponding dimension tables: sample,marker,assay,term

Gene_fact is liked to its corresponding dimension tables: gene, go, pm_id, dm_id, cl_id

Experiment_fact is linked to its corresponding dimenson tables: norm, protocal, person, project, publication, platform.

Microarray_fact is linked to its corresponding dimension tables probe, measure unit

All these links are through their respective primary keys of the individual dimension tables.

Furthermore, the dimension tables are also linked to one another as: clinical_fact , experimental_fact and gene_fact are linked to microarray_fact with foreign keys.

The OLAP layer was implemented on top of the Oracle DB in Java.

While doing a time complexity analysis, we find out that typical queries entered by the user onto the oracle db takes O(n) or O(logn) depending on weather it is a full table scan or an indexed scan.

Implementing an OLAP layer adds an additional amount of complexity.

i.e for Queries 1,2,3 in part II it adds a performance penalty ofO(n). However asymptotically the performance remains O(n) as O(n)/O(logn) + O(n) remains O(n) in asymptotic notation.

For query 4,5 the performance penalty is O(n) + O(n) as we need to find the mean and variance.

For query 6 the performance penalty is an additional O(n^2) as the correlation every element in one list must be found against the other.

For Part III, The performance penalty is an additional O(n^2) performance penalty imposed on the olap layer.

Query Results:

Part II

Query 1:The number of patients who have tumor are 53,The number of patients who have leukemia are 27 and the number of patients who have ALL are 13

Query 2:There are a total of 19 types of drugs which have been applied to patients with tumor.

Query 3:The number of mRNA values in cluster ID 0002 and patients with ALL and measure unit 001 is 325.

Query 4:The t Statistics of the expression values between patients with ALL and patients without ALL and for probes belonging to go ID 0012502 is -1.007126.

Query 5:For probes belonging to go ID 0007154,the f statistic of the expression values among patients with ALL, AML, colon tumor, breast tumor is 3.1389.

Query 6:For probes belonging to go ID 0007154 the average correlation of the expression values between 2 patients with ALL is 0.14354.

The average correlation of expresiion values between 1 ALL patient and ! AML patient is -0.0034756.

PartIII

Query 1: For patients with ALL serving as control group A and not ALL serving as control group B, the number of informative genes is 41.

Query2:Given a new patient ,test 2 is classified into the group which has ALL and test 1,test3,test 4,test 5 are classified into the group which has NOT ALL.