CSE601

DATA_MINING_PROJECT1

Prepared By:

Anuj Rastogi (50134324)

Nalin Kumar (50170479)

Pranshu Pancholi (50169864)

Part1

Data Warehouse Implementation, Design, Performance and Improvements

In the initial stages, we imported the dataset provided to us in separate text files into different tables in our database in the exact same manner as in given files. Subsequently, we performed some data cleaning steps such as converted all nulls appearing in all id columns to an integer value of -99 to get a much more clear picture so as to clearly distinguish nulls and non-null values and also to maintain and keep similar data types in every individual columns. Considering the dataset size, we understood that star constellation best fits the picture when the dataset size is limited and as the dataset size grows bigger, something similar to biostar model best fits that particular kind of scenario. In our case, when we directly dumped the data into database, we saw that there too many columns and too many null values in most of the fact tables. Based on that, initially we figured out that although according to the dataset size, star constellation could have been directly used to answer all the problems, but still we went ahead and implemented something similar to biostar schema model since we realised that as the dataset size gets bigger and bigger in the future, we might need to add many more columns in the fact tables and the null values in those tables will only get amplified with time. Henceforth, keeping in mind the bigger picture, we went ahead with something close to biostar schema model since we believed that this would help us considerably in scaling our database in the future as more data gets added up. Biostar type of schema also helps in normalizing our tables, since we broke down the fact tables into many small tables, thereby reducing the amount of redundant data in the fact tables which can be a big problem in the future. Although, one disadvantage of biostar schema is that since we added up many intermediate tables between dimension tables and the main fact tabes which can be called as link tables, we needed to perform more joins while answering the queries to the problems. But the many advantages listed above clearly offsets this single disadvantage of biostar schema model since it many advantages in terms of scalability, flexibility, reliability and extensibility along with keeping minimal amount of redundant data and null values in most of the tables, especially the main central fact tables.

When we analysed the performance of our data warehouse schema, we realised that most of the operations occurred in approximately n^2 time complexity (n is the data set size on which query is getting executed) since we needed to perform many joins in answering some complex problems

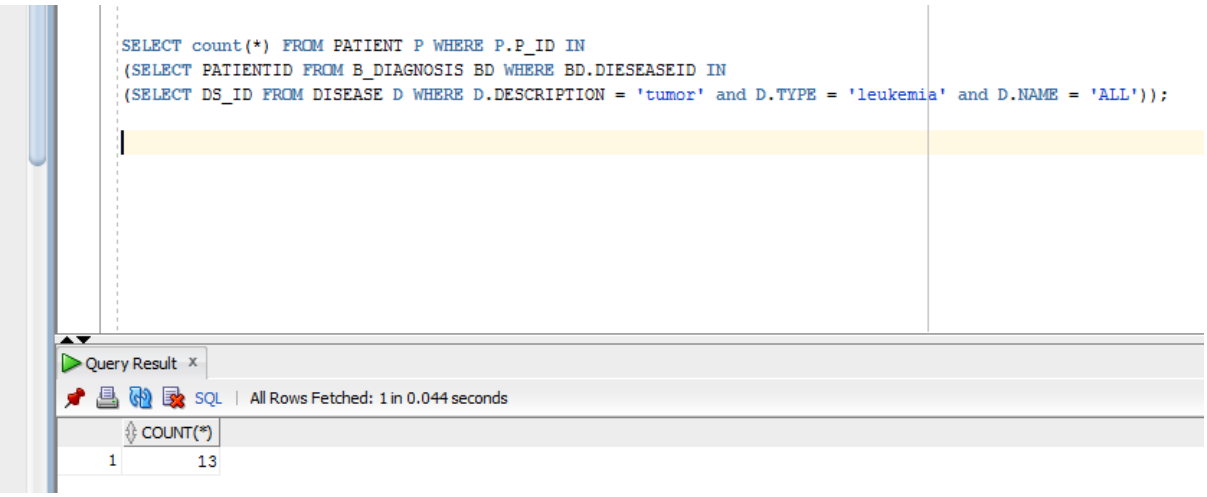
which reduced our performance. But as the database grows larger and larger this reduced performance overhead will surely be overtaken by the simplicity and scalability of our data warehouse design. In order to improve the performance of our database in the future, we can add indexes on specific columns in most of the tables on which queries are executed most frequently. This might improve the performance of our design in the future and it might take close to logarithmic time in some of the operations since database indexing on columns uses a B-Tree kind of structure which logarithmic time complexity. In the end, we would like to state that our database is highly capable of supporting many kinds of OLAP operations such as roll-up, drill down, slice etc. and few of these appear in the problems explained in the sections below.

Part2

In the first part, we designed our data warehouse in such a manner that it supports regular OLAP operations as well as some statistical operations. For answering the specific parts of this problem, we used SQL queries to answer the first 5 problems and wrote code in Java using JDBC connection to solve the last problem involving evaluating average correlation. The explanation for the various parts in this problem is described below.

Problem 2.1

In this part we were supposed to list the number of patients whose disease description is tumor, disease type is leukemia and disease name is all. We answered this part using a simple query which evaluates all those patients whose disease id corresponds to the above mentioned details. There are in total 13 patients which matched the specified problem statement.The query and its result snapshot is as follows:-



Problem 2.2

In this part we found the type of drugs which have been applied to all the patients having disease description as tumor. There are in total 220 drug types as depicted in the following query and results snapshot.

```
SELECT TYPE FROM DRUG D WHERE D.DR_ID IN
(SELECT DRUGID FROM B_DRUGUSE BD WHERE BD.DRUGID NOT IN (-99) AND BD.PATIENTID IN
(SELECT PATIENTID FROM B_DIAGNOSIS BD WHERE BD.DIESEASEID IN
(SELECT DS_ID FROM DISEASE D WHERE D.DESCRIPTION = 'tumor')));
```

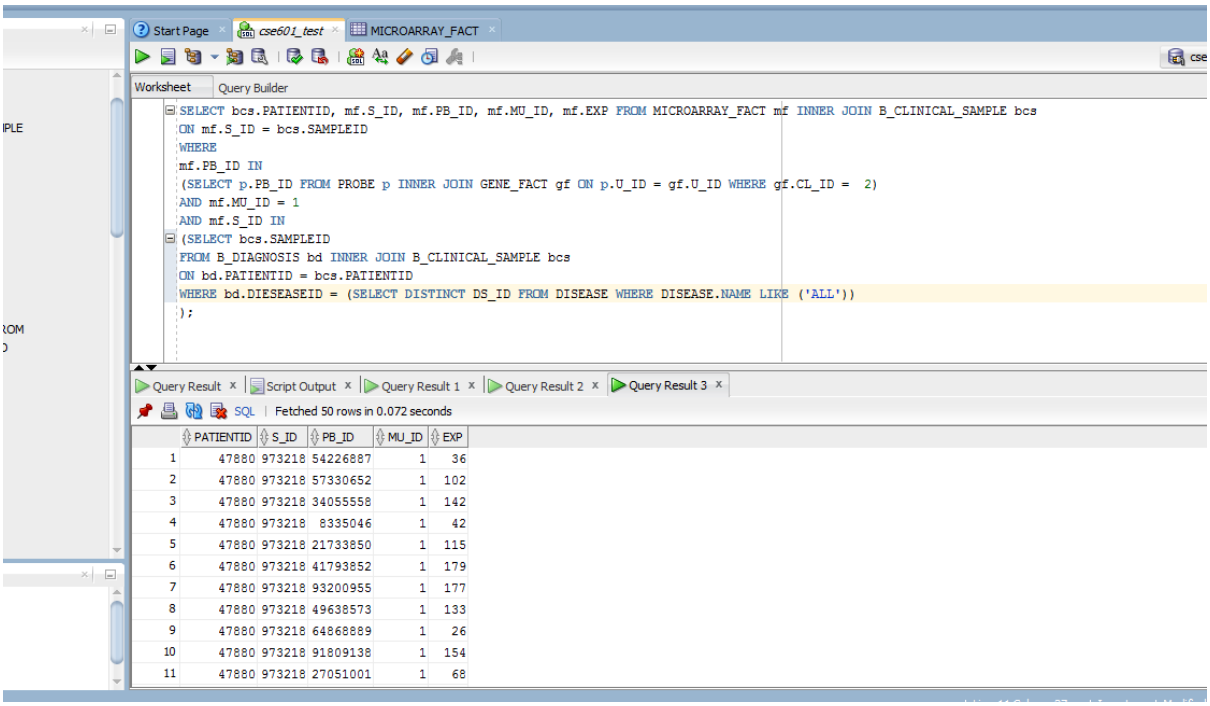
Query Result

SQL | All Rows Fetched: 220 in 0.244 seconds

	TYPE
210	Drug Type 007
211	Drug Type 003
212	Drug Type 005
213	Drug Type 017
214	Drug Type 004
215	Drug Type 016
216	Drug Type 009
217	Drug Type 014
218	Drug Type 013
219	Drug Type 016
220	Drug Type 012

Problem 2.3

In this part, we were asked to list the mRNA values (expression) of probes in cluster id = 2 for each experiment with measure unit id =1 for each sample of patients with all. The query and results snapshot for this part are as follows:-



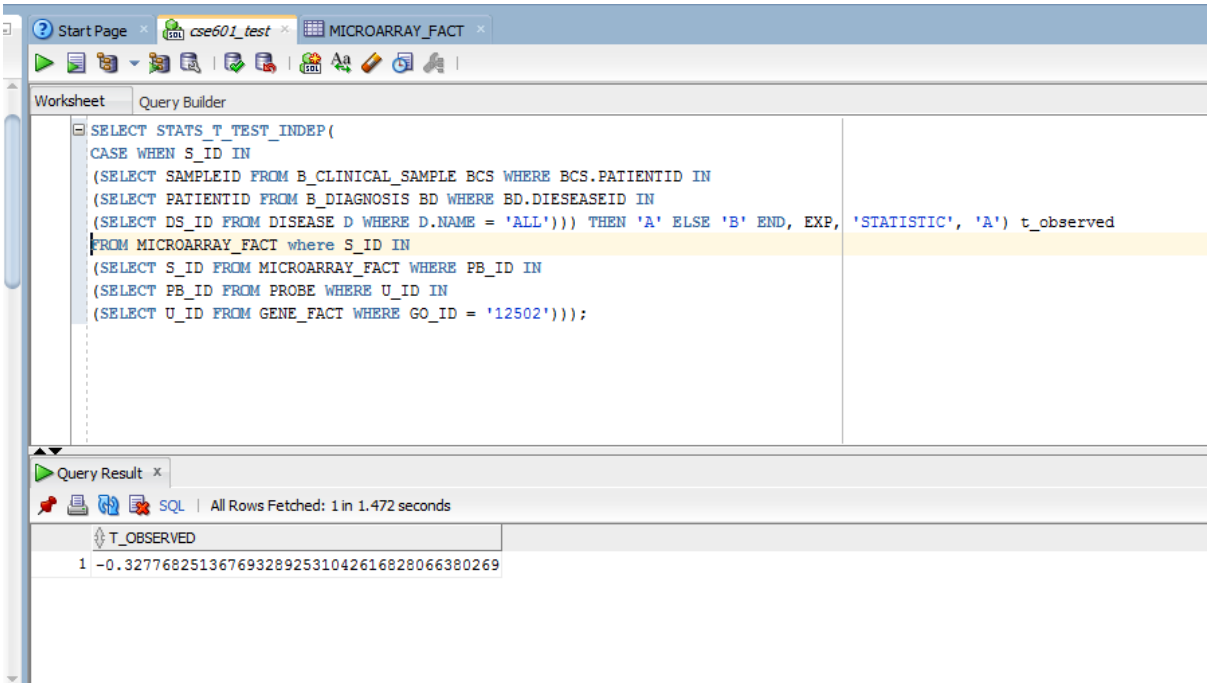
```
SELECT bcs.PATIENTID, mf.S_ID, mf.PB_ID, mf.MU_ID, mf.EXP FROM MICROARRAY_FACT mf INNER JOIN B_CLINICAL_SAMPLE bcs
ON mf.S_ID = bcs.SAMPLEID
WHERE
mf.PB_ID IN
(SELECT p.PB_ID FROM PROBE p INNER JOIN GENE_FACT gf ON p.U_ID = gf.U_ID WHERE gf.CL_ID =  2)
AND mf.MU_ID = 1
AND mf.S_ID IN
(SELECT bcs.SAMPLEID
FROM B_DIAGNOSIS bd INNER JOIN B_CLINICAL_SAMPLE bcs
ON bd.PATIENTID = bcs.PATIENTID
WHERE bd.DIESEASEID = (SELECT DISTINCT DS_ID FROM DISEASE WHERE DISEASE.NAME LIKE ('ALL'))
);
```

Fetched 50 rows in 0.072 seconds

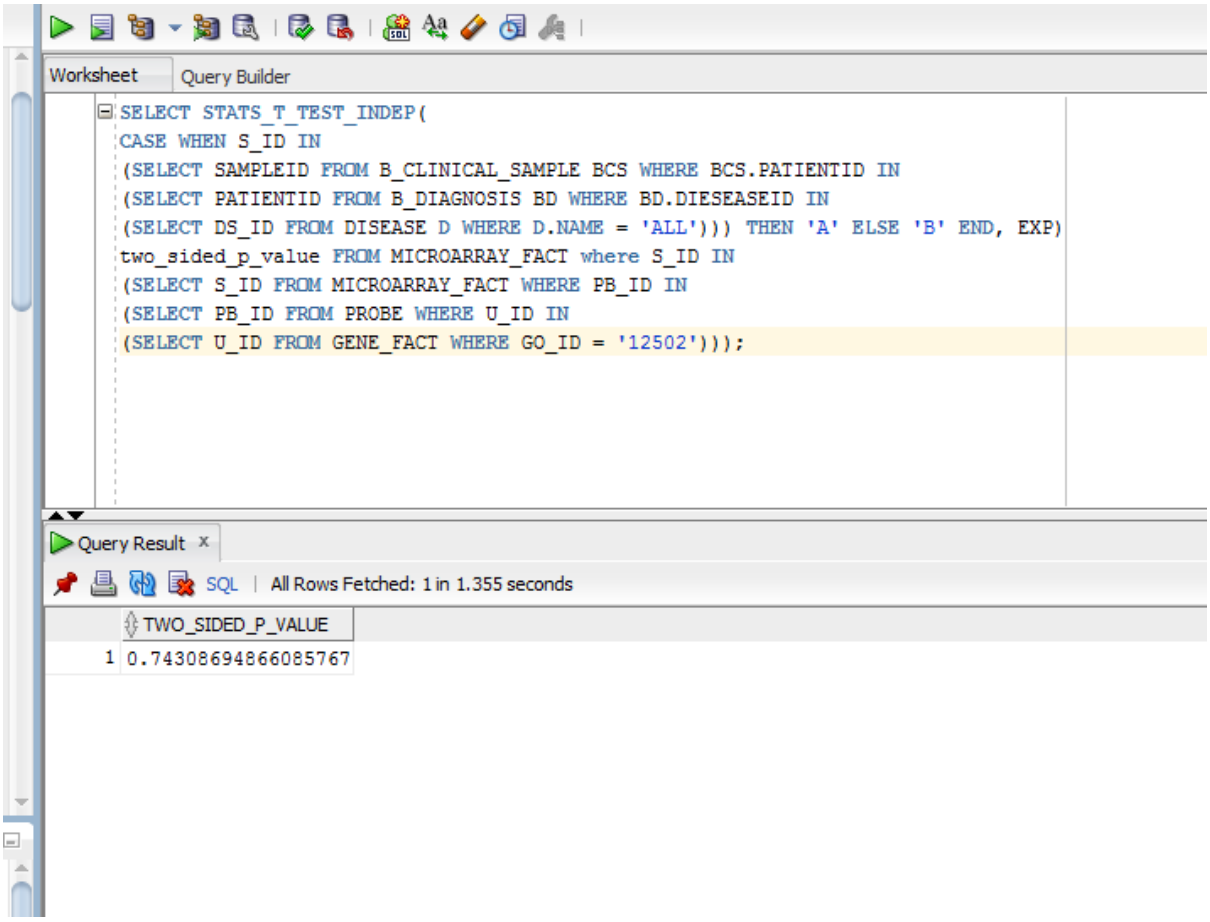
	PATIENTID	S_ID	PB_ID	MU_ID	EXP
1	47880	973218	54226887	1	36
2	47880	973218	57330652	1	102
3	47880	973218	34055558	1	142
4	47880	973218	8335046	1	42
5	47880	973218	21733850	1	115
6	47880	973218	41793852	1	179
7	47880	973218	93200955	1	177
8	47880	973218	49638573	1	133
9	47880	973218	64868889	1	26
10	47880	973218	91809138	1	154
11	47880	973218	27051001	1	68

Problem 2.4

In this part, we were supposed to calculate the t statistics of the expression values between patients with all and without all for probes belonging to go id = 12502. The t test measures the significance of difference of means of two samples. Here, our two samples are patients with all and patients without all. We used STATS_T_TEST_INDEP Oracle built in function since it was given that we need to calculate t statistics for two samples under the assumption that they are having equal variance. We used this same function two times differently, once to calculate t observed and again to evaluate p value associated with the t value. For this, we included the first parameter as a categorical variable which has value ‘A’ for patients with all and value ‘B’ for patients without all and the second parameter as expression value of a particular gene in a particular patient. The query and results for calculating the t values are as follows:



The query and results snapshot for p values are as follows:-



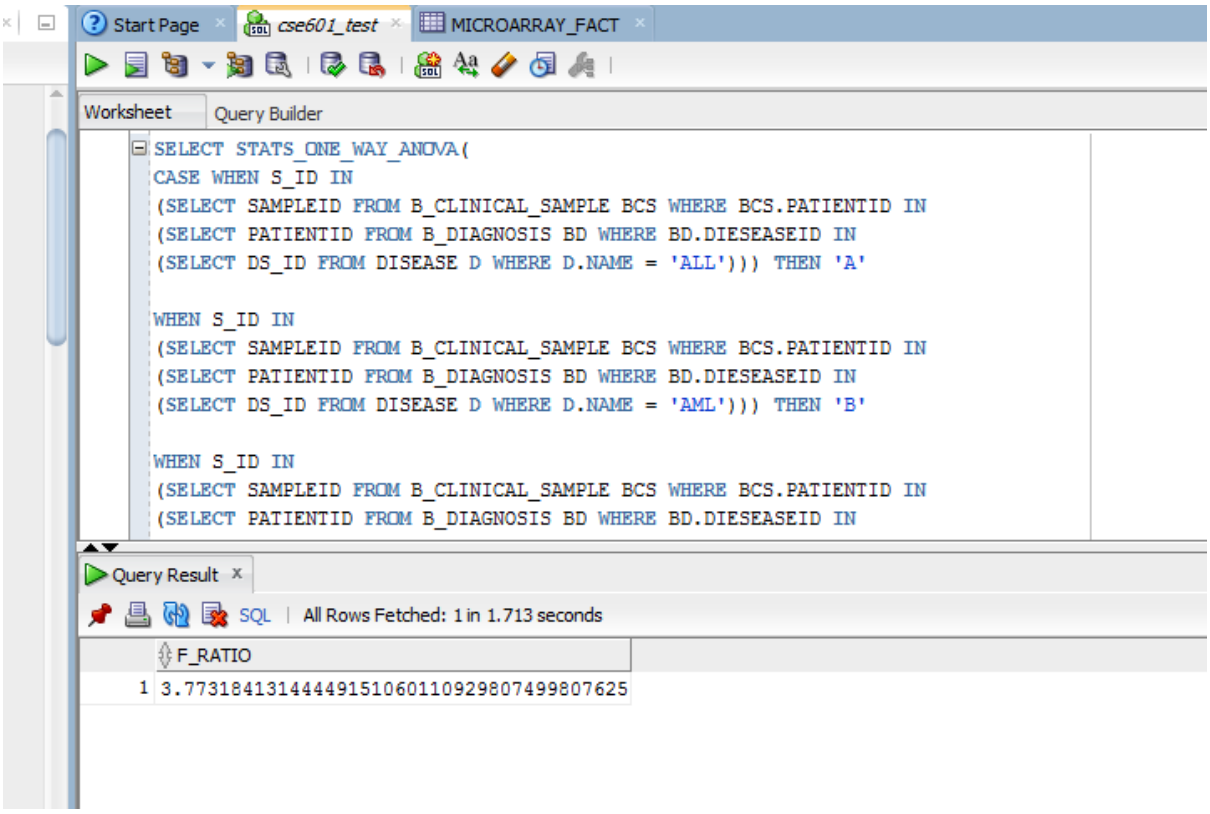
As can be seen from the above images, the t value came out to be -0.327 and the p value corresponding to the above t value came out to be 0.743. When we are performing a t test, we are essentially trying to find the evidence of a significant difference between 2 population means. The greater the magnitude of T (either positive or negative), the greater the evidence against the null hypothesis that stats that there is no significant difference between the two samples. The closer T is to 0, the more likely is the fact that there isn't a significant difference between the two samples. The p value of a t test is very closely linked to its t value since the larger the absolute value of the t value, the smaller the p value and the greater the evidence against the null hypothesis.

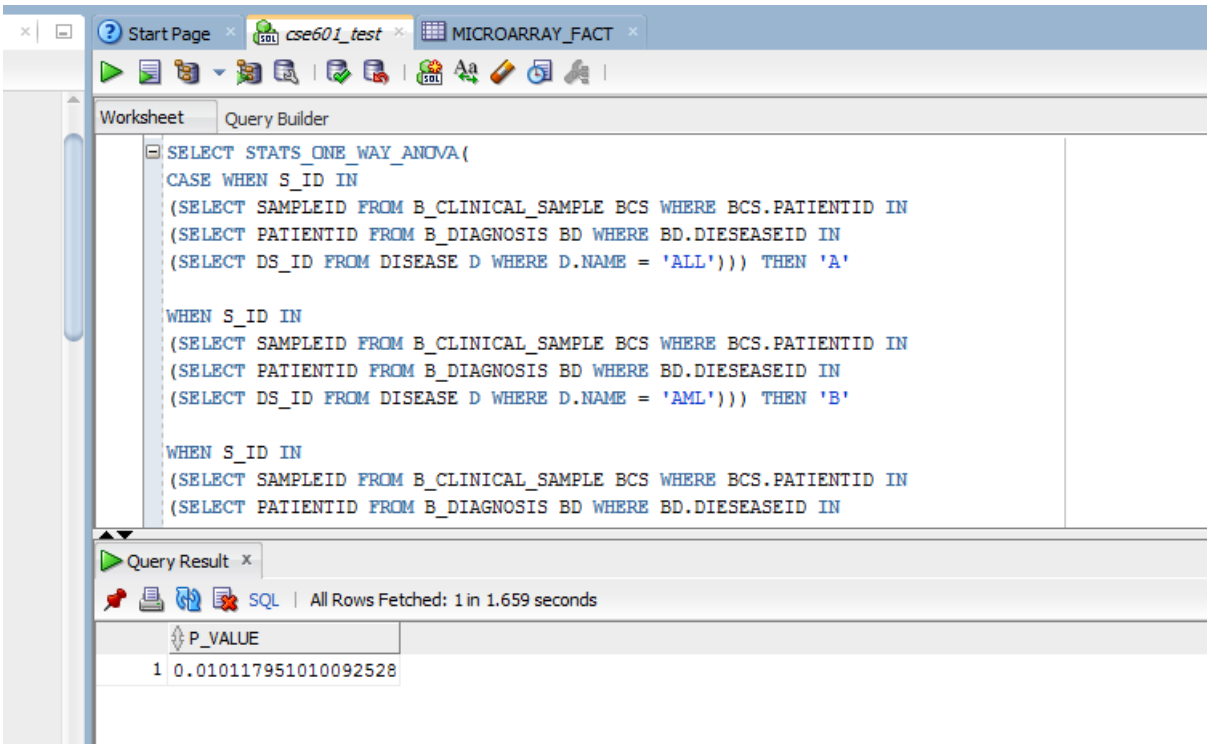
In our case, since the t value is close to zero and the p value is on a higher side (much greater than the significance level, assuming that alpha = 0.05 which means that there is a high risk of

going wrong against the null hypothesis), we can claim that the null hypothesis holds true which essentially means that there isn't a significant difference between the two samples of patient with all and patient without all with probes belonging to go id = 12502.

Problem 2.5

In this part, we need to evaluate the f ratio and its corresponding p value to calculate the F statistics of the expression values among patients with all, aml, colon tumor and breast tumor. We evaluated the f ratio and p value of the F statistics using the same STATS_ONE_WAY_ANOVA Oracle function with just a bit of difference in the number of parameters specified. Analysis of variance (ANOVA) can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means. The first snapshot below calculates the f ratio amongst the 4 groups and the second snapshot calculates the p value associated with the f ratio amongst the 4 groups of disease name categories for probes belonging to go id = 7154. As in previous part, the first parameter is a categorical variable and takes values A, B, C and D depending on whether the disease name is all, aml, colon tumor or breast tumor. The query and results snapshot for both f ratio and p value are as follows:-


Start Page
cse601_test
MICROARRAY_FACT
Worksheet
Query Builder
SELECT STATS_ONE_WAY_ANOVA(
CASE WHEN S_ID IN
(SELECT SAMPLEID FROM B_CLINICAL_SAMPLE BCS WHERE BCS.PATIENTID IN
(SELECT PATIENTID FROM B_DIAGNOSIS BD WHERE BD.DIESEASEID IN
(SELECT DS_ID FROM DISEASE D WHERE D.NAME = 'ALL'))) THEN 'A'
WHEN S_ID IN
(SELECT SAMPLEID FROM B_CLINICAL_SAMPLE BCS WHERE BCS.PATIENTID IN
(SELECT PATIENTID FROM B_DIAGNOSIS BD WHERE BD.DIESEASEID IN
(SELECT DS_ID FROM DISEASE D WHERE D.NAME = 'AML'))) THEN 'B'
WHEN S_ID IN
(SELECT SAMPLEID FROM B_CLINICAL_SAMPLE BCS WHERE BCS.PATIENTID IN
(SELECT PATIENTID FROM B_DIAGNOSIS BD WHERE BD.DIESEASEID IN
Query Result
SQL
All Rows Fetched: 1 in 1.713 seconds
F_RATIO
1
3.773184131444491510601109298074998076 25



As can be seen above, the F ratio value came out to be 3.77 and the corresponding p value came out to be 0.010. F test actually determines whether group means are equal and is a measure of the ratio of variation between sample means as numerator and variation within individual samples as the denominator. Henceforth, in order to show that the sample or group means are different, we need to show that the sample means are further apart from each other and within the group variance is low. F statistic is a ratio of the two quantities explained above that are expected to be roughly equal under the null hypothesis, which produces an F-statistic of approximately 1. The lower the F value, the closer the group means are to each other relative to the variability within each group. The higher the F value, the larger is the variability of group means relative to the variability within group. In our case, The F value is on a higher side but we also need to scrutinize upon p value to reach any conclusion since F values can also be different if we pick multiple random samples of the same size from the same population. The p value is the probability that allows us to determine how common or rare our F value is under the assumption that the null hypothesis is true. If the p value is low enough, we can conclude that the given data is inconsistent with the null hypothesis. In our case, the p value came out to be slightly more than 0.01 which is low enough to reject the null hypothesis using the common significance level of

0.05. Henceforth, based on F ratio and the corresponding p value, we can conclude that the patients categorized into 4 different groups in the problem have more variation amongst their group means relative to the variability within each group and hence they are different.

Problem 2.6

In this problem, we are asked to evaluate the average correlation of the expression value between two patients with all and the average correlation of the expression values between one all patient and one aml patient. Since this problem required to calculate multiple pairwise individual correlations for all combinations of patients in both the case, we evaluated the average correlation by writing code in Java using JDBC connection to connect to our data warehouse schema perform certain queries before arriving at the final result.

Pearson's correlation is used to assess the strength and direction of association between two variables that are linearly related to each other. Its coefficient, r, indicates the strength and direction of this relationship and can range from -1 for a perfect negative linear relationship to +1 for a perfect positive linear relationship. However, a value of 0 indicates that there is no relationship between the two variables. In our case, as can be seen from the output of the Java program (Part2Question6.java), the results came out to be as follows:-

```
11. Key= 70835, Value=[49.0, 30.0, 171.0, 104.0, 114.0, 26.0, 181.0, 10.0, 164.0, 42.0,
12. Key= 17659, Value=[49.0, 30.0, 171.0, 104.0, 114.0, 26.0, 181.0, 10.0, 164.0, 42.0,
13. Key= 304, Value=[49.0, 30.0, 171.0, 104.0, 114.0, 26.0, 181.0, 10.0, 164.0, 42.0, 2
14. Key= 48802, Value=[49.0, 30.0, 171.0, 104.0, 114.0, 26.0, 181.0, 10.0, 164.0, 42.0,


Correlation Sum Amongst All Patients: 78.0
Average Correlation Amongst ALL Patients: 1.0

Correlation Sum Amongst ALL and AML Patients: 30.271064381960734
Average Correlation Amongst ALL and AML Patients: 0.16632452957121283
```

As can be seen from the output of the program, there were in total 13 patients with all and 14 patients with aml for probes belonging to go id = 7154. We took all the pairwise correlations into consideration before calculating the final average correlations. It can be clearly seen that there is a perfect positive linear relation amongst the expression values of the patients belonging to all with go id = 7154. However, in the other case, the average correlation value came out to be significantly lower, that is 0.166, which means there is a week linear relationship amongst the expression values of all and aml patients which can be logically deduced as since the patients belonged to different disease name categories, their week linear relationship of expression values indicates that there is a significant difference in gene expression values for these two different disease name categories.

PART 3

In this part, we had 2 subparts. Hence, the explanation below is divided into the following.

PART# 3.1

This part was focused mainly on finding the informative genes. Before doing so there is the first question which could be answered directly by using SQL. All the remaining part for this subpart require JAVA code. Execution rules for this are already provided in the README.txt file provided.

3.1.1

We need to find ALL patients in Group "A" (ALL), Group "B" (REST).

The following query was used to get the results.

```
-----Question 3.1.1--------------ALL PATIENTS CLASSIFIED AS DIFFERENT GROUPS-------------------

SELECT DISTINCT(PATIENTID),
CASE WHEN DIESEASEID = 2 THEN 'A' ELSE 'B' END AS GRP
FROM B_DIAGNOSIS WHERE DIESEASEID NOT IN (-99);
```

A snapshot of the obtained results is also attached. This is just the top few rows of the result.

	PATIENTID	GRP
1	79777	B
2	92978	B
3	95052	B
4	53880	B
5	70863	A
6	6413	B
7	31076	B
8	22162	A
9	6060	A
10	68707	B
11	99163	B
12	86986	B
13	52573	B
14	16521	B
15	28582	B
16	47360	A
17	90893	B

3.1.2/3.1.3

When we combine these two parts, we mainly need to find the informative Genes, which can help us do prediction on a new patient. The genes with P-Value less than 0.01 are considered

For this part we need to use the JAVA Code provided. The instructions on how to use JAVA Code are given in the README.txt file. Once the JAVA Code is. executed we will get the T-test values for all Genes, and also the P-Value for all the genes. Also, the result provides a list of informative genes along with their P-Values. All the results will be output to a text file as per the implementation of the program.

The following JAVA Code file has to be used. Part3Question1.java. In our case, the file generated has been provided. **The name is "informativeGenes.txt".** The file provides a T-test and P-Value for each gene. Then in the lower part of the file there is a list of informative genes with their P-Values.

Here is the list of informative genes we found out with their P-Values:

```
1. Gene UID: 4826120 P Value: 2.3530152E-8
2. Gene UID: 83398521 P Value: 1.2779025E-6
3. Gene UID: 40567338 P Value: 1.7602703E-7
4. Gene UID: 37998407 P Value: 0.008551328
5. Gene UID: 43866587 P Value: 2.817193E-8
6. Gene UID: 13947282 P Value: 0.007978611
7. Gene UID: 31308500 P Value: 1.0766322E-5
8. Gene UID: 58792011 P Value: 4.12953E-6
9. Gene UID: 74496827 P Value: 8.066191E-9
10. Gene UID: 85557586 P Value: 2.534077E-6
11. Gene UID: 60661836 P Value: 4.1949384E-6
12. Gene UID: 41333415 P Value: 2.949359E-7
13. Gene UID: 48199244 P Value: 0.008580554
14. Gene UID: 88257558 P Value: 6.0011894E-7
15. Gene UID: 15295292 P Value: 1.6939097E-6
16. Gene UID: 21633757 P Value: 7.7214545E-6
17. Gene UID: 58672549 P Value: 7.2289918E-6
18. Gene UID: 69156037 P Value: 1.9798094E-7
19. Gene UID: 53478188 P Value: 2.996341E-8
20. Gene UID: 97606543 P Value: 5.446245E-4
21. Gene UID: 41464216 P Value: 3.0292108E-7
22. Gene UID: 88596261 P Value: 1.21376315E-5
23. Gene UID: 94113401 P Value: 0.0076199076
24. Gene UID: 18493181 P Value: 4.1190597E-6
```

```
25. Gene UID: 45926811 P Value: 0.007301182
26. Gene UID: 11333636 P Value: 0.0034437159
27. Gene UID: 1433276 P Value: 1.2585128E-8
28. Gene UID: 31997186 P Value: 0.0012435749
29. Gene UID: 28863379 P Value: 2.9730666E-6
30. Gene UID: 47276861 P Value: 0.00979001
31. Gene UID: 52948490 P Value: 5.5619416E-6
32. Gene UID: 75434512 P Value: 0.005772966
33. Gene UID: 92443312 P Value: 0.005364478
34. Gene UID: 24984526 P Value: 0.009262996
35. Gene UID: 75492172 P Value: 3.9120386E-7
36. Gene UID: 16073088 P Value: 4.353061E-7
37. Gene UID: 87592194 P Value: 0.008080939
38. Gene UID: 65772884 P Value: 4.7083915E-9
```

Explanation of the logic:

Basically when we calculate the T-Test value corresponding to a Gene-ID, what we do is we first find out the expression values for Gene-ID, in one group and then in other group. Now, using the

STATS_T_TEST_INDEP function from oracle we find out the T-Test value. T-Test values mainly, tell us if the difference between the means of the two sets, can be generalized to the population as a whole.

If the value of T-Test is large, then we can generalize, results from sample to whole population, saying that the Gene-ID's, expression values have significant difference on grouping of patients among "ALL" and "Rest of diseases". Also, if the T-test value is large the p-value is small. This indicates that with how much confidence can we say that this gene is the once whose expression values create a difference.

*In our case we have chosen the P-value as 0.01. That means that, if a confidence level of 99.99% is there that the expression values corresponding to this gene is creating the difference, then we regard this gene as Informative gene. Hence, based on the above we have found out **38** informative genes.*

PART# 3.2

In this part we were given a list of patients and we had to find out if they belonged to "ALL" or NOT, based on the informative genes found out in the above part. The question had several subparts.

3.2.1

For the first part the following script was executed in SQL.

```
--------QUESTION 3.2.1-----------FINDING INFORMATIVE GENES WRT ALL(GROUP A)------------------------------
--------------------How informative genes have been founded is done as a part of the problem in Part3 1,2,3--------
SELECT * FROM GENE WHERE U_ID IN (
SELECT DISTINCT(p.U_ID)
FROM B_DIAGNOSIS bd
INNER JOIN B_CLINICAL_SAMPLE bcs ON bd.PATIENTID = bcs.PATIENTID
INNER JOIN MICROARRAY_FACT mf ON bcs.SAMPLEID = mf.S_ID
INNER JOIN PROBE p ON mf.PB_ID = p.PB_ID
WHERE bd.DIESEASEID IN (2) AND p.U_ID IN (4826120,
83398521,40567338, 37998407, 43866587, 13947282,31308500,
58792011,74496827,85557586,60661836,41333415,48199244,88257558,
15295292,21633757,58672549,69156037,53478188,97606543,41464216,
88596261,94113401,18493181,45926811,11333636,1433276,31997186,28863379,
47276861,52948490,75434512,92443312,24984526,75492172,16073088,87592194,65772884)
);
```

The list of Informative genes that we got from the upper part was given as an input. A sample snapshot is given below to reflect the kind of output we got.

	U_ID	SEQTYPE	ACCESSION	VERSION	SEQDATASET	SEPCIESID	STATUS
1	58792011	Gene Type 091	Hs.93316	4.5	NCBI	8560	G
2	18493181	Gene Type 049	Hs.92351	3.5	NCBI	8348	K
3	65772884	Gene Type 019	Hs.55432	2.8	NCBI	5844	J
4	4826120	Gene Type 026	Hs.51326	1.9	NCBI	8117	I
5	21633757	Gene Type 035	Hs.67107	3.8	NCBI	6351	U
6	41464216	Gene Type 035	Hs.77418	1.2	NCBI	9808	U
7	31308500	Gene Type 042	Hs.59898	6.2	NCBI	3460	L
8	74496827	Gene Type 030	Hs.90784	3.5	NCBI	4004	C
9	97606543	Gene Type 031	Hs.26697	5.3	NCBI	233	I
10	88596261	Gene Type 049	Hs.4605	3.9	NCBI	6558	J
11	31997186	Gene Type 043	Hs.87960	8.7	NCBI	6049	U
12	24984526	Gene Type 027	Hs.453	8.9	NCBI	40	F
13	75492172	Gene Type 019	Hs.70549	7.3	NCBI	2585	F
14	1433276	Gene Type 077	Hs.67352	2.8	NCBI	5718	G
15	88257558	Gene Type 008	Hs.82525	4.8	NCBI	6871	G
16	52948490	Gene Type 089	Hs.86276	5.8	NCBI	7377	K
17	43866587	Gene Type 030	Hs.17093	7.8	NCBI	7929	M
18	16073088	Gene Type 006	Hs.67405	9.1	NCBI	2670	O
19	83398521	Gene Type 064	Hs.31915	0.7	NCBI	2540	X
20	92443312	Gene Type 008	Hs.91316	3.8	NCBI	1737	D
21	75434512	Gene Type 069	Hs.82101	4.8	NCBI	569	J

3.2.2

Find all the patients with ALL.

For this part the following SQL script was executed.

```
-----QUESTION 3.2.2---------Find all Patient with ALL group A-----------------------------------

SELECT PATIENT.P_ID,PATIENT.GENDER,PATIENT.NAME,PATIENT.SSN,bd.DIESEASEID
FROM B_DIAGNOSIS bd INNER JOIN PATIENT ON bd.PATIENTID = PATIENT.P_ID WHERE DIESEASEID = 2;

-------------------------------------------------------------------------------------------------
```

The results are as follows:

	P_ID	GENDER	NAME	SSN	DIESEASEID
1	47880	Male	Eiapvj Sseory	827-14-4546	2
2	13258	Male	Btxuda Eldhmg	731-97-2100	2
3	79175	Female	Cvhlur Dnvwxg	800-62-2970	2
4	33553	Female	Fgbdsp Kbbwhq	503-92-3261	2
5	70863	Female	Uyendk Qaunsq	602-99-2042	2
6	765	Male	Dvsdyi Aqcbxb	226-07-9236	2
7	65736	Male	Lanvdo Iszyjq	717-27-2128	2
8	6060	Female	Dpchyq Ionwwz	365-01-2585	2
9	58484	Male	Slnvyz Gdqgqu	364-35-8439	2
10	47360	Male	Qlcnmz Aabsme	226-37-2462	2
11	2378	Female	Ynegrr Bppjnn	608-50-2104	2
12	22162	Male	Jrgibn Bcxeix	750-88-0515	2
13	77689	Male	Rzxnci Bdjxsf	443-65-4129	2

3.2.3/3.2.4/3.2.5/3.2.6

For these parts, we have a JAVA code, “Part3Question2.java” that needs to be executed. The instructions on how to execute that code are given in the README.txt file. Once executed, it gives results on the output screen classifying the patients as either “ALL” or “NOT”, based on the informative genes found in the above part.

Also, we needed to calculate the co-relation coefficients of new patient, with each patient from group “ALL” and with each patient from group “NOT ALL”. Based on these co-relation values we will use the T-Test to determine if the Patient belongs to “ALL” or “NOT ALL”.

The following results were obtained when we ran it for the conditions provided in the project.

```
Coefficients of Co-relation:

Patient 1 and Group ALL Patients

[0.7399504281406035, -0.13315969625393892, 0.8599105864877905, -
0.05695069681635744, 0.8028348873995631, 0.8087932463330986,
0.8362402033547757, 0.017460239849087705, 0.7632604816394647,
0.7650925013667067, 0.81552101443019, 0.7912667293386927, 0.8060622904244571]

Patient 2 and Group ALL Patients

[0.24013715477262237, -0.09157422189380415, 0.20824274359743927,
0.295431909001944, 0.17362912131652752, 0.24039991739456185,
0.11476333210383319, -0.050465120001738396, 0.261132961086563,
```

0.16171323825540726, 0.1273346074189273, 0.16604639795440823,
0.21150667919497]

Patient 3 and Group ALL Patients

[-0.04849729483826874, -0.3836804015628954, 0.054517184680760375,
0.17495241500709927, -0.16599821455994912, -0.011924268321006655, -
0.016327282280226764, 0.06452008744949955, -0.03528459157046645, -
0.04974383778424124, 0.06184563629143548, 0.008067350815596898, -
0.09023607599317587]

Patient 4 and Group ALL Patients

[0.7213662402847247, -0.05357598004090899, 0.8641239289548529,
0.06442748923613534, 0.8102129205741408, 0.8058409141089488,
0.7918513096454549, -0.0284976363441692, 0.7512167189032196,
0.7088269379178248, 0.8637512071323489, 0.8255837928819985,
0.8097460483397702]

Patient 5 and Group ALL Patients

[-0.16265885250496073, 0.0807216643517056, -0.12553010154056296, -
0.21092631502890255, -0.13955197108579995, -0.09493302652913534, -
0.11909672841099765, -0.10845309386782631, -0.15997174861321511, -
0.18254323427380198, -0.11254378217964034, -0.1336297057274836, -
0.12178852527684611]

Patient 1 and Group NOT ALL Patients

[-0.10566003566048582, -0.28700868901452203, 0.07703291650938966,
0.09245815332191751, 0.16616275680953738, -0.25643454595395143, -
0.0217779073989387, 0.11868954891016484, 0.2642950022688474,
0.07722875696045836, 0.30902907633504173, 0.12855760162053317, -
0.0542290803335932, 0.234371355684041, 0.3300465991269529,
0.30097836013568224, 0.15373105522120323, -0.37754960243990965, -
0.008818963579056156, 0.01893090558969417, -0.15675720225315706, -
0.1079110412149225, -0.10774200031094883, 0.12085790499339204, -
0.1554228273631989, 0.25150248797157204, 0.19646963664642722,
0.2861397926149318, 0.03821992815393601, -0.14080210347302838,
0.06893544557362359, 0.15562883534300803, 0.10143700044648434,
0.1336124390220071, -0.03929429025720936, 0.10051811588798557, -
0.1904569030941274, -0.15678493760703596, 0.04192555852606254, -
0.07326030804488713]

Patient 2 and Group NOT ALL Patients

[0.07898822176534735, -0.22117583210598682, 0.012135544927207774,
0.08477900607955317, -0.1555455559741812, 0.13915552199676495, -
0.006318104754945173, -0.27748482033511745, 0.11535319144437096,
0.04316917306119809, 0.0613084710287907, -0.27322731840042336, -
0.14451051778305282, 0.13039962183163667, 0.0360055773238499, -
0.11267374059201637, 0.11510031910657659, 0.0014783447823465056, -
0.014421638948549587, -0.13722482481921866, -0.1252305629722604, -

0.25039070663661006, 0.06888770734516465, 0.09574090655653698, -0.040406501682569484, 0.10791254064938409, -0.21857907041643054, -0.03517431747139833, -0.05765716949731178, -0.2546682480051865, -0.16037438081481434, -0.14128074387088366, -0.016854659190564395, 0.05336382384371078, -0.04194561294899645, -0.19239096992263494, -0.16357058503411373, 0.08309066119965725, -0.0334855594394494, -0.08384380485553948]

Patient 3 and Group NOT ALL Patients

[0.16743195518160547, 0.1846834928859359, 0.12148332205026782, -0.07905965398748675, -0.16997299270684385, -0.16709328448946575, 0.06661196865845481, 0.014549823527618702, -0.24675114726098013, -0.14138978534969351, -0.1661426416957377, -0.1375207814394341, 0.0967936405473, -0.008507267222115542, -0.09302094399443814, -0.03598386926220725, -0.04358584009797578, 0.2192277575710607, 0.03967742944900807, 0.05364199242205009, 0.028894880412707002, -0.13806659698795007, -0.08185045197387705, 0.1917841980822115, -0.10726948931199139, -0.06631857284114365, -0.08372782646456814, 0.04838220581027303, 0.16942262952621784, -0.10350204301913579, -0.06372300329478614, 0.04011823556563349, -0.07619201251115461, 0.011808356994954434, -0.024620224489279693, 0.0467288179861271, -0.011866505133034475, 0.16407512737132363, 0.049520010996515006, 0.039713798257675896]

Patient 4 and Group NOT ALL Patients

[-0.046531729319488256, -0.1807896634253015, -0.03173530978683108, 0.08322310075668066, 0.06336747146100115, -0.18005863389767335, -0.017080406039412264, 0.015007370019841735, 0.27120120087951144, 0.09352989941298753, 0.321034139439802, 0.1009337049474745, 0.0412183246675858, 0.19876141204576397, 0.1919072921974084, 0.2329290963414326, 0.09608723383456114, -0.35895909585028435, 0.011014431668758625, -0.0595330157216668, -0.06567272057534773, -0.21191695222777562, -0.12961015470881707, 0.1462631963760399, -0.0802436439003031, 0.2694688583120036, 0.016484361608419845, 0.2777487749255009, 8.445733226579718E-4, -0.08183620026554728, 0.07123848485179023, 0.20621022362195962, 0.11224219244221119, 0.10252942627718502, 0.008052096385826782, 0.12262785366084669, -0.1438836325489768, -0.05605967069002029, 0.020291850848248427, -0.06620143610159382]

Patient 5 and Group NOT ALL Patients

[0.3136890136011124, 0.16600967613811105, -0.1449862470659601, 0.004960394158375931, -0.2026453684084287, 0.1525068402596411, -0.20484881221830117, -0.4001640378561057, -0.036097909150070634, -0.16685283822685076, 0.026625895644938922, 0.10950698908645023, -0.1658158647335773, 0.13363036935105554, 0.10930633364758642, -0.2374837480769431, -0.07379638115902643, 0.12606154882592438, 0.07817443163770887, -0.19503508700317182, 0.1623010555558183, 0.09298429860433544, -0.3113060011793777, -0.02128113191068409, -0.09291991464827742, 0.2159462396662039, -0.1352302002243487, -0.2711793830412984, 0.04204763974313026, -0.11142183265307326,

```
0.0016445211624207747, 0.22859486270312462, 0.12959651199964833,
0.22599495789002536, 0.11836842629405275, -0.04060799069284, -
0.18275446351095082, 0.12996927160914126, 0.20945389697211894,
0.08811098854329241]
```

Also, the Classification comes out to be:

```
T-Test
Patient 1 : is : ALL :  P-value:  1.3707148152162773E-9
Patient 2 : is : ALL :  P-value:  3.0538147966540447E-6
Patient 3 : is : NOT ALL :  P-value:  0.49384947004821034
Patient 4 : is : ALL :  P-value:  4.1167568040764866E-11
Patient 5 : is : NOT ALL :  P-value:  0.019572619804261688
```

Hence, based on the above results, when analyzed on informative genes, these are the results for patients.

EXPLANATION

Once, we found out the informative genes we can say with surety that these are the genes which matter when it comes to a patient belonging to ALL and NOT ALL. Hence, these are the genes whose combination of expression values defines if the patient will belong to ALL or NOT.

Now, based on these informative genes, we calculated the coefficient of correlation of expression values between a new patient and all the other patients of ALL group. Also, we find a coefficient of correlation between expression values between a new patient and other patients of NOT ALL group.

The when we find the t-test between the coefficients of correlation, it tells us if the correlation of the patient with one group is significantly different from the other and based on that and the P-Value we predict if the new patient belongs to ALL or NOT. Here we have used P-Value as 0.01, which means that only when we have a confidence of more than 99.99% are we going to say that correlation are different and hence, patient belongs to ALL.