MSA 8040 Data Management for Analytics: Project 1

Houping Xiao, Georgia State University September 15, 2022

Requirements (Please read it carefully!):

(a) Please keep it confidential and do not distribute it outside our cohort!

(b) Report and Demo must be given for each group.

(c) Report should be submitted on iCollege before 11:59 pm on10/09/2022.

(d) This is group-based project. Only one group member need to submit the files.

(e) Individual-based peer evaluation form. Each group member need to submit their own evaluation for the other group members in your group.

Peer-Evaluation Form

To better achieve fairness in the class, at the end of the course you will be asked to evaluate yourself and the other members of your group on completing the project. These ratings are used for gauging team members' contributions. The grade you and your group members receive will depend in part on these peer evaluations. Rate each member based on the following criteria: (1) participation in group activities, (2) quality of work, (3) quantity of work, (4) finishing assigned work on time, and (5) ability to work as a team member. Please use the following scale to assign scores:

5	Exceptional effort, above and beyond the call of duty
4	Above average effort
3	Normal effort (this is the expected score!)
2	Below average effort
1	Unacceptable effort

Then, submit the following note to the instructor on iCollege:

Your Name:_____ Score:_____ Reasons:_____

Team Member #2:_____ Score:_____ Reasons:_____

Team Member #3:_____ Score:_____ Reasons:_____

Team Member #4:_____ Score:_____ Reasons:_____

Team Member #5:_____ Score: _____Reasons:_____

Note: Please include a brief reason for any group member. I expect everyone to be thoughtful and diligent in completing this evaluation. Your final project grade will be biased by the pear-evaluation scores fro your group members. For instance, you could be graded as ZERO for the project if you receive "1"s from all other group members.

1 Database

In this project, you are asked to implement a clinical and genomic database based on your schema design using the MySQL system. A good database should satisfy the following requirements: (1) support regular queries and statistical OLAP operations; (2) be robust to potential changes in the future; an (3) support knowledge discovery.

The original data will be provided in the plain text files **Project1.zip**. A detailed description of the file format is attached at the end. The information related to MySQL system can be found at:

- https://www.mysqltutorial.org/

The queries statement can also be found in the lectures Notes on the website

- https://houpingx.github.io/database.html

1.1 Part I [50%]:

You are required to design a database schema [**20%**]; implement your database schema in the MySQL system and populate your database with the provided data sets (**30%**).

1.2 Part II [50%]:

Your database is supposed to support SQL operations and regular OLAP operations. In the following are some typical queries by users. You may use either SQL, PL/SQL or external programs (e.g., in Java) to answers the queries. Notice that you should retrieve the data from the MySQL system instead of the original plain text files. Report your approach and the results returned by your database.

- List the number of patients who had “tumor” (disease description) [**5%**], “leukemia” (disease type) [**5%**] and “ALL’ (disease name) [**5%**], separately.
- List the types of drugs which have been applied to patients with “tumor” [**5%**].
- For each sample of patients with “ALL”, list the mRNA values (expression) of probes in cluster id “00002” for each experimennt with measure unit id=“001”. [**10%**] (**Note:** measure unit id corresponds to mu_id in microarray_fact.txt, cluster id corresponds to cl_id in gene_fact.txt, mRNA expression value corresponds to exp in microarray_fact.txt, UID in porbe.txt is a foreign key referring to gene_fact.txt.)
- For probes belonging to GO with id = “0012502”, calculate the difference between the average of the expression values of patients with “ALL” and the average of expression values of patients without “ALL”. [**10%**]
- For probes belonging to GO with id=“0007154”, calculate the average of the expression values among patients with “ALL”, “AML”, “colon tumor” and “breast tumor”, and order by the average value in a descending order. [**10%**]

A Appendix: Descriptions of data file format

For each files in **Project1.zip**, it is considered as an entity which starts with a row describing the fields of the entity. Then, each following row in the file corresponds to one instance of the entity.

A.1 Clinical data space

Entities: patient, disease, drug, test and sample; Fact table: clinical_fact.

- patient.txt (p_id, ssn, name, gender,DOB)
- disease.txt (ds_id, name, type, description)
- drug.txt (dr_id, name, type, description)
- test.txt (tt_id, name, type, setting)
- clinical_fact.txt (p_id, ds_id, sympton, ds_from, ds_to, dr_id, dosage, dr_from, dr_to, tt_id, result, tt_date, s_id)

A.2 Sample data space

Entities: sample, marker, assay, term; Fact table: sample_fact.

- sample.txt (s_id, source, amount, sp_date)
- marker (mk_id, name, type, locus, description)
- assay.txt (as_id, name, type, setting, description)
- term.txt (tm_id, name, type, setting)
- sample_fact.txt (s_id, mk_id, mk_result, mk_date, as_id, as_result, as_date, tm_id, tm_description)

A.3 Microarray and proteomic data space

Entities: probe, measureUnit; Fact table: microarray_fact.

- probe.txt (pb_id, UID, name, description, isQC)
- measureUnit.txt (mu_id, name, type, description)
- microarray_fact.txt (s_id, e_id, pb_id, mu_id, expression)

A.4 Gene data space

Entites: gene, go, cluster, domain, promoter; Fact table: gene_fact.

- gene.txt (UID, seqType, accession, version, seqDataset, speciesID, status)
- go.txt (go_id, accession, type, name, definition)
- cluster.txt (cl_id, num, pattern, tool, tSetting, description)
- domain.txt (dm_id, type, db, accession, title, length, description)
- promoter.txt (pm_id, type, sequence, length, description)
- gene_fact.txt (UID, go_id, cl_id, dm_id, pm_id, UID2)

A.5 Experiment data space

Entities: experiment, project, platform, norm, person, protocal, publication; Fact table: experiment_fact.

- experiment.txt (e_id, name, type)
- project.txt (pj_id, name, investigator, description)
- platform.txt (pf_id, hardware, software, settings, description)
- norm.txt (nm_id, type, software, parameters, description)
- person.txt (pn_id, name, labName, contact)
- protocal.txt (pt_id, name, text, createdBy)
- publication.txt (pu_id, pub_med_id, title, authors, abstract, pubDate)
- experiment_fact.txt (e_id, nm_id, pj_id, pn_id, pf_id, pt_id, pu_id)