Modelling Database Requirements with ER Diagrams

Matriculation Number : 220032472

06th October 2022

Contents

Task 1

1.1 Intended use of the system

The system developed in this scenario is a university lab which performs the following actions:

1. Storing research projects and samples information that are being analysed in the lab.
2. To track information of samples provided by each researcher in research projects.
3. To analyse and store results of each sample analysed in a batch.
4. Research project related sampling information is queried from the system.
5. Sending researchers notification/reminders to collect the samples after analyses.
6. Revenue statistics generated by analysing samples for the research projects.
7. To formulate the efficiency and demand of the work done at the lab.

1.2 User of this system

1. University – To get overall account of the lab including service statistics to analyse demand and efficiency of the machines.
2. Student and staffs (or) Users – Use the machine to analyse and get the report of their sample.
3. Finance team – To get the cost/revenue generated.

1.3 Queries

1. Which entity in the ER model indicate that a batch is suitable for being processed by a machine based on the specifications provided for it.
2. The requirement states that a machine can manage multiple samples in a batch, so will the batch entity hold multiple samples (Our machines have varied characteristics, including maximal number of samples).
3. Which all entities will provide information on billing of samples that are analysed for a month – Finance related information.
4. From which entity will you get the data for sending notification to the researcher once the results are generated or if the sample remains in storage for longer period?
5. How will you check whether a staff member or a student has the credentials to charge its analysis to this project?
6. What happens on reanalysis of a sample?
7. How can we get statistical data of the work?

1.4 Assumption

1. Every batch is sent to the machine for analyses
2. Every machine gives a result for all samples in a batch – no downtime, or irregularities in its analysis.
3. Every project has a researcher, and every researcher or user must be a part of project

4. Only one user must provide every sample, i.e., no two users can own a sample.
5. Every reanalysis is added to a new batch – as the sample unique id and weight will always be the same to keep track of multiple analysis of same sample. But the batch in which the analysis was done must be different , so that the result will also be a separately saved value.

1.5 Specification

1.5.1 Entity Specifications

1. **Machine** (name, batch_runtime, max_batch_size, max_weight_of_samples, model, year_of_manufacturing)
   The Machine schema includes a 'name' as a unique identifier and contains information related to the batch_runtime and max_weight_of_samples which are positive values along with the year of manufacturing of that machine and its model.

2. **Batch** (batch_id, processing_start_datetime, number_of_samples(), total_weight_of_samples())

   The schema Batch contains 'batch_id' as its unique identifier.
   processing_start_datetime indicates when the batch was sent to machine for analysis.
   The derived attributes in this case are the number_of_samples() and total_weight_of_samples().

   The derived attribute total_weight_of_samples(), can be used to decide the machine, as each machine have a limit on the maximum weight of samples.

   The derived attributes number_of_samples() can be used to compute the efficiency of the machines.

3. **Sample** (sample_id, weight)
   The schema 'Sample' has 'sample_id' as its unique identifier and the domain of the attribute 'weight' must be a positive value for each sample.

4. **User** (email, name, department)
   It contains two subclasses
   a. **Staff** (job_title)
   b. **Student** (start_date)

   The 'User' schema is a superset of staff and students, and it contains email, name and department for each user as its attributes, where 'email' becomes the unique identifier as no two users can have same email address.

   The subclass 'staff' has additional attribute named 'job_title' and the entity 'student' has an additional attribute named 'start_date'.

5. **Project** (sort_code, account_number, name)
   The 'Project' schema refers to the research project that each user is part of. It contains information pertaining to the account information such as 'sort_code' and 'account_number' acting as a unique composite attribute.

6. **Result** (analysis_end_datetime, {result_of_analysis}, sample_delivered_on_datetime, age_for_notification())
   The 'Result' schema is a weak entity set and it relies on batch entity set via the relationship 'saves'. The primary key of the entity is 'analysis_end_datetime' and combines with 'batch_id' from batch entity.

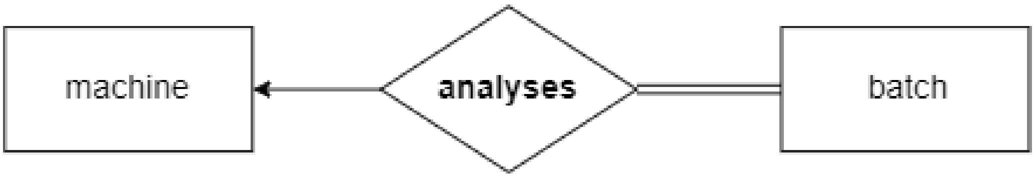
   The result entries are created when a machine completes an analysis of a sample in a batch and it stores the result of the analysis in a CSV file associated with a particular sample of a batch as attribute names as {result_of_analysis}

   sample_delivered_on_datetime indicates the date and time of when the user collected the sample after analysis.

   Attribute age_for_notification() is derived from the analysis_end_datetime to post a notification to the researcher if the samples are not collected within 10 days after being analysed.

1.5.2 The relations in the University lab E-R diagram

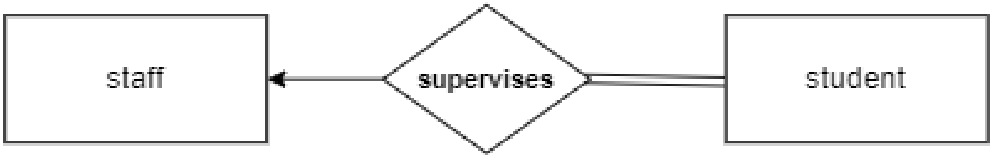
1. **analyses**



The entity sets 'machine' and 'batch' are related via a relationship set 'analyses'.
Mapping cardinality constraint: Many-to-One – An entity in batch is associated with at most one entity in machine. An entity in machine, however, can be associated with any number (zero or more) of entities in batch.

Participation constraint:
The participation of an entity set 'batch' in a relationship set 'analyses' is **Total** as every entity in 'batch' participates in at least one relationship in 'analyses.
The participation of an entity set 'machine' in a relationship set 'analyses' is **Partial** as only some entity in 'machine' participates in relationships in 'analyses' .

2. supervises



The ‘student’ and ‘staff’ are subclasses of entity set ‘user’. The ‘student’ entity set is supervised by ‘staff’, where every student must have a supervisor.

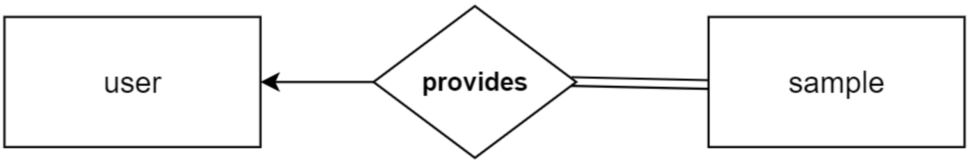
Mapping cardinality constraint: One-to-Many – An entity in student is associated with at most one entity in ‘staff’. An entity in ‘staff’, however, can be associated with any number (zero or more) of entities in ‘student’.

Participation constraint:

The participation of an entity set ‘student’ in a relationship set ‘supervises’ is **Total** as every entity in ‘student’ participates in at least one relationship in ‘supervises’, i.e., each student much be supervised by a staff member.

The participation of an entity set ‘staff in a relationship set ‘supervises’ is **Partial** as only some entity in ‘staff’ participates in relationships in ‘supervises’, i.e., not all staff need not supervise students.

3. provides



The relationship between ‘user’ and ‘sample’ is via relationship set ‘provides’ where each sample must be given by a ‘user’. And each ‘user’ can provide multiple samples, and each sample can be provided by only one user.

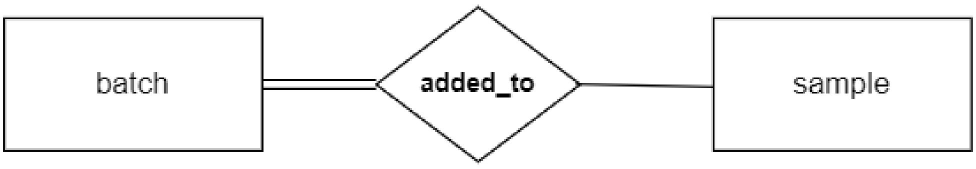
Mapping cardinality constraint: One-to-Many – An entity in ‘sample’ is associated with at most one entity in ‘user’. An entity in ‘user’, however, can be associated with any number (zero or more) of entities in ‘samples, stating that a user can submit any number of samples.

Participation constraint:

The participation of an entity set ‘sample’ in a relationship set ‘provides’ is **Total** as every entity in ‘sample’ participates in at least one relationship in ‘provides’, i.e., each sample must be given by a user.

The participation of an entity set ‘user’ in a relationship set ‘provides’ is **Partial** as only some entity in ‘user’ participates in relationships in ‘provides’, i.e., some user need not provide samples, but can still access the system to get previous results.

4. added_to

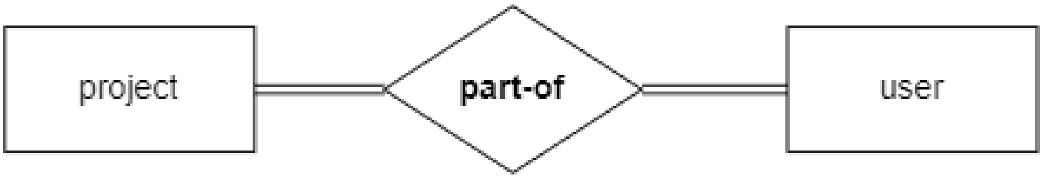


The relationship between ‘batch’ and ‘sample’ is via relationship set ‘added_to’. The samples added to the batch are then passed on to machine for analysis.

Mapping cardinality constraint: Many-to-Many – An entity in ‘sample’ is associated with zero or more entities in ‘batch’. An entity in ‘batch’ is also associated with any zero or more number of entities in ‘samples’, stating that a user can submit any number of samples, and there can be multiple batches of samples from users.

Participation constraint:
The participation of an entity set ‘batch’ in a relationship set ‘added_to’ is **Total** as every entity in ‘batch’ participates in at least one relationship in ‘provides’, i.e., each batch must be created from the provided samples only.
The participation of an entity set ‘sample’ in a relationship set ‘added_to’ is **Partial** as only some entity in ‘sample’ participates in relationships in ‘added_to’, i.e., some sample need not be added to part of batch.

5. part_of

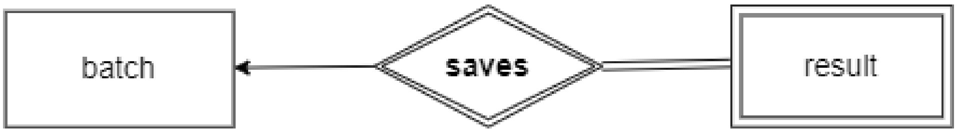


The relationship between ‘user’ and ‘project’ is via relationship set ‘part_of’.

Mapping cardinality constraint: Many-to-Many – An entity in ‘user’ is associated with zero or more entity in ‘project’. An entity in ‘project’ are associated with any number (zero or more) of entities in ‘user’, stating that a user can submit any number of samples, and a project can have multiple users providing samples.

Participation constraint:
The participation of an entity set ‘user’ in a relationship set ‘part_of’ is **Total** as every entity in ‘user’ participates in at least one relationship in ‘part_of’, as the system creates a record only when a user as part of a research project brings in a sample.
The participation of an entity set ‘project’ in a relationship set ‘part_of’ is **Total** as at least one entity in ‘project’ participates in relationships in ‘part_of’, i.e., each project must have one or more users associated with it.

6. saves

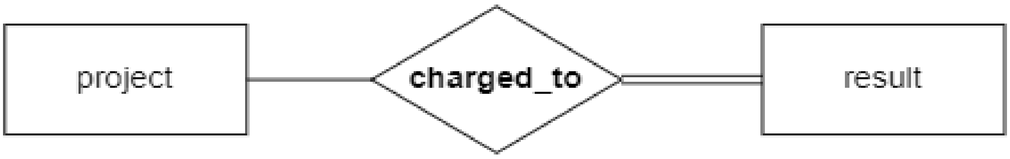


The relationship set ‘saves’ is used to indicate the reports generated by each batch to be saved in result entity set. The result entity set will not contain a unique id of its own, however it is dependent on batch for its existence, which is made explicit by making it a weak entity set. The discriminator of a weak entity set is underlined with a dashed line.

The relationship set between result and batch is indicated with a double diamond indicating a relation between a strong and a weak entity.

Weak entity in this case the ‘result’ always has total participation but strong entity which is ’batch’ may not have total participation.

7. charged_to



The relationship between ‘result’ and ‘project’ is via relationship set ‘charged_to’.

Mapping cardinality constraint: Many-to-Many – An entity in ‘result’ is associated with one or more entity in ‘project’. An entity in ‘project’ are associated with any number (zero or more) of entities in ‘result’, stating that a sample that is part of a research project can be re analysed by producing separate result set for every analysis.

Participation constraint:

The participation of an entity set ‘result’ in a relationship set ‘charged_to’ is **Total** as every entity in ‘result’ participates in at least one relationship in ‘charged_to’, as every analysis are billed to a research project.

The participation of an entity set ‘project’ in a relationship set ‘charged_to’ is **Partial**.

Task 2

2.1 Entity-Relation diagram for University Lab

Design a representation of the data in terms of entities, attributes and relationships between entities. Construct an **E-R diagram** to depict this representation.

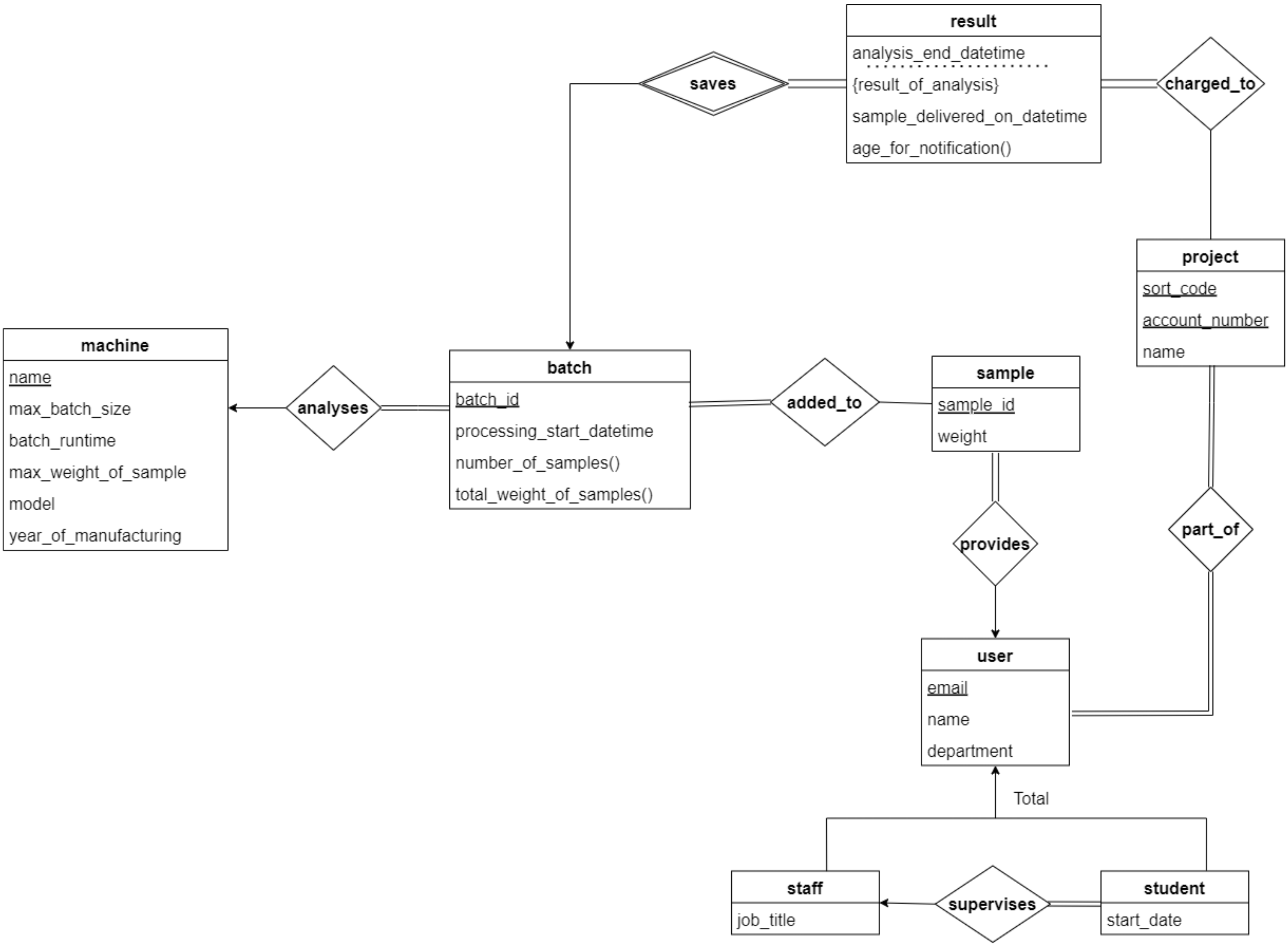


Fig 2.1: E-R diagram for a university lab.

2.2 Linking ER model to queries in Task 1

Q1. Which entity in the ER model indicate that a batch is suitable for being processed by a machine based on the specifications provided for it.

> The batch entity in the ER model stores the information of the samples that are added as part of that batch that needs to be analysed in a machine. All samples are weighed when they are received by the university lab. This allows to compute the total weight of the batch before being fed to the machine. The information regarding total number of samples and total weight of all the samples can be derived in the batch entity and can be used in determining the machine that can be selected for analysis.

Q2. The requirement states that a machine can manage multiple samples in a batch, so will the batch entity hold multiple samples (Our machines have varied characteristics, including maximal number of samples).

The information about the samples characteristics is passed to machines via batch entity set. Entity Batch holds all the information need to map a group of samples to a machine that meets the specification and maximum weight and size constraints that a machine has.

Q3. Which all entities will provide information on billing of samples that are analysed for a month – Finance related information.

The result entity set stores the information about the samples and when the researcher collected them after analysis. If the analysis_end_datetime is within this month then cost can be added to the corresponding months finance report.

It also stores the results of all the analysis made under a batch, this way the total number of sample and their weight is also known, which can be utilized in computing the cost incurred on a research project.

Q4. From which entity will you get the data for sending notification to the researcher once the results are generated / How will the time for sending notification be calculated.

The result entity set records information on when the analysis of a sample was done, and when the sample was picked up by the user from the lab storage. A derived attributed named age_for_notification() can be computed if an analysis was completed ,but the samples were not picked up from the lab, i.e., if the sample_delivered_on_datetime does not exists. If the age exceeds 10 days from the date of analysis completion, then a reminder will be sent to researcher.

Q5. How will you check whether a staff member or a student has the credentials to charge its analysis to this project?

The user/researchers are mapped to project, so when a sample is being submitted for analysis a check on this entity and its relationship with entity set project can determine if a user has the credentials to charge its analysis to a project.

Q6. What happens on re-analysis of a sample?

The sample with the same unique_id and weight that needs to be re analysed are added to a batch and a corresponding result for the new analysis is produced which can be billed to a project. Since the result entity set is weak and depends on the information of the batch entity such as processing start date, batch_id and analysis_end_datetime in result can act as a identifier for finding information related to this new analysis for same sample.

Q7. How can we get statistical data of the work?

The demand of the machine can be derived from the batches that are added to machine for analysis and the result that each machine produced for the corresponding samples in a batch. The date time attributes in batch and result can be used to check the demand of a machine.

The efficiency can be computed from the number of batched analysed by each machine over a period, as we capture the time taken for each analysis for each batch of sample.

Task 3

3.1 Understanding the importance of ER model

Creating a system and understanding how we can build an entire system without complexities and allowing for scaling in the future is understood from the ER modelling. By having a diagram that clearly shows the entities and relationships that are representing, it is much more obvious to identify the areas that are related. It is also easier to visualize the impact of a change in one part of the model on other parts of the model.

Visually representing a system with its entities and relationships opens design phase for stakeholders to get involved in phrasing questions that compliments in building a system that is less prone to errors or loss of information.

Entity Relationship model provided an intuitive way to communicate the database design.

3.2 Challenges faced during modelling the ER diagram

- The kind of assumptions to be made before modelling the system. This was resolved by assuming a real-life scenario of the system and how the operations of a university lab would look like on a day-to-day basis. This formed a clearer picture of the system that needed to be designed.
- Modelling the system using extended ER concepts were challenging. This was resolved by referencing individual concepts under the extended ER modelling concepts in lecture 3 and 4 from university of St. Andrews and applying them to the ER model created using the specification and assumptions, to see where a refactoring could be done for bettering the system.

3.3 Problems faced during modelling and its resolution

One of the biggest problems that was encountered during the modelling process was redundant attributes. This was analysed to be due to lack of complete information/self-understanding on basic ER concepts and not having a solid grasp of the extended ER features.

This was tackled by referencing the lecture notes, cross verifying with the textbooks mentioned in the lecture slides for further reading and using sample ER models developed for complex systems.

The reasons for looking at complex systems was because they had resolved the common issues that can occur when you do an ER modelling such as the Fan trap and utilized standardization/generalization wherever needed. This gave an idea of a clean model which are free from redundant attributed and entities.

The theory sessions were time bound and varieties of problem statements could not be analysed for a given concepts. The content of the textbooks also seemed limited to a specific example that are analysed for explaining a concepts. I would like to come up with few samples or real time scenarios that encompass large spectrum of systems in order to explain the varied possibilities of cases that can be associated to a concept of modelling ER diagrams.

3.4 Short reflective session

- The designing process made me think in perspective of a stakeholder, providing me with questions that could arise when developing a software system.

- Without a good set of questions, a clean ER model cannot be built. As lot of refactoring was based out of these questions. The mapping, standardization/generalization concepts were possible only when referencing to the questions that were formulated before the modelling process.

- The complexities that arise due to improper understanding of the basic DBMS concepts related to ER modelling needs to be avoided by preparing a well-structured note for each concept before diving into the modelling process.

- Had to look up at multiple resources to conclude good practises and standards rather than referring to single source. There were discrepancies when referring to multiple online web pages. A reliable source of information needed to be analysed and followed throughout the coursework to maintain the consistency of design.

  The reference books mentioned in the lecture slides were of a good help when forming the constraints and was to be used as a single source for most part of the coursework.

- With respect to ER model, I was convinced towards the end that a good ER model can save up time during a system development process, as there were many instances recorded by organizations that they had to redo a lot of development towards the end of the software lifecycle as their initial ER modelling was not catering to most of the requirements. This was indicated to have costed companies a hefty sum of money. So, with a good ER modelling by considering all scenarios, the software can be developed and shipped to the end user in a smooth and clean manner. This information gave a good insight on the importance of this coursework and how it needs to be done with precision in the real-world applications.

References

[1] Database System Concepts by Abraham Silberschatz (Yale University), Henry F. Korth (Lehigh University), S. Sudarshan (Indian Institute of Technology, Bombay), sixth and seventh edition, Published by McGraw-Hill Education

[2] IS5102 – Database Management Systems, Lecture Notes, University of St Andrews, 2022 - 23.

[3] Introduction to ER Model - Geeks for Geeks website (https://www.geeksforgeeks.org/introduction-of-er-model).