**FA23: DATA-225 Sec 11 - Db Systems for Analytics**

**Homework - - 1**

**Name :- Prayag Nikul Purani**

**SJSU Id :- 017416737**

**Question 1a:-**

1. Binary file :-

The binary files are the low-level files which are used by the system as the computer files to store the data in form of zero’s and one’s or the binary format. The data stored in these files are only readable by the machine and it difficult for a human being to decode it. So, the question arises that why this file are used over our normal language because we untimely be the one who will be using this files. The answer to the above riddle is that this file are efficient and compact to store data than normal text file. This type of files are generally created by the compilers, which translate the high level human language to low level machine – readable language. The have extension of .bin or .exe. There is no metadata available for this types file. As it is hard to decode it so the modification that can be made is also possible. As this types of files are compact and efficient so are used to store videos, images, bitmaps, and audio files.

Advantages:-

* These files normally are compact and use less storage.
* They are faster to compile and execute.

Disadvantages:-

* Very complex to understand.
* Not easy to edit.
* These files require a software to run.

1. Text file :-

The text file are the most basic file used to save textual data containing alphabets, number, symbols, notations and so on. Unlike binary file this types of file are written by the end user which are easy to make, edit and modify depending on the situations. Apart from this features there are more option to enhance the data such bold, italics, underline and the users can also insert images to make their work easier. The other name for this files are flat files or ASCII files. In windows operating system the text files are created using text editors like Notepad or Microsoft word and it has the extension of .txt or .docx. On the other hand this text editors can to write source code of different programing language such JAVA, HTML, PHP ad etc. we just have to change the extension according to the usability of the user. They can be written or read on simple text editors. The main disadvantages of this files are data redundancy and the confidentially of the data is reduce. While storing large amounts of data, text files can not be an ideal tool to use as it becomes tiresome and dull for a user to keep all the records and information in one place.

Advantages:-

* Easy to understand
* Modification and readability increases.

Disadvantages:-

* These files require more space.
* Execution is slow.

1. Text file with key value pair :-

These types of files are as normal text files but the data is stored in more organized, structured and systematic manner. The data stored in this files are easy to read, edit and processing of the data is also bit easy in compare the text files. The structure of this files are like dictionary in terms of our programing language. The parsing in this types of file are bit easy. It generally requires more space for storage than binary files. Apart this the main drawback of this type is scalability the efficiently and the risk to duplication increase with the increase in the quantity of the data. This type of file are use in XML, JSON, CSS and etc., depending upon the usage the extension of the file depends.

Advantages:-

* Easy to implement and read/write.

Disadvantages:-

* Execution is slow
* It is hard to add any new record because we need to declare key again and again to add new data.

1. Excel Spreadsheet :-

The Excel spreadsheet was developed by Microsoft. It is use to store data in tables in form of rows and columns. Each value is stored in a cell and the notation of it in form of (row, column) like F8 is the 8th cell in the 'F' column. It can store and maintain large amounts of data. It has many in-built functions which are used for editing and formatting the data. It has many inbuilt features such as sum, average and many mathematical operations could be done on the data, also there are many graphs option with is a great help to a data analytics to do his job more quickly. Task such as data management, data analysis, accounting, charting and graphing etc can be done on excel spreadsheets. Excel spreadsheet include built in functions and formulas for better analysis.

Advantages:-

* Data input is very easy
* Can be transferred to different formatted such csv and so on, also vice versa is also possible.
* Multiple spreadsheet can be formed and access at same time.
* Good for projects for shorter duration.

Disadvantages:-

* Learning and implementation of excel functions and formulas is quite difficult.
* Cannot handle large data

1. Relational Database:-

A relational database is a collection of data that are related and have some all-ready defined relationship. It is highly structured and organized as tables with rows and columns. Each entity in the logic from real world is stored in tables with its features called attributes, also different entity can be connected using relationship between them. A system is required to control relational database called database management system (DBMS). It can store all types of data and maintains ACID properties. Data stored in the database can be accessed and modified with the help of SQL queries. It provides the highest level of security, consistency, and integrity. In case of a server crash, it recovers the data with the help of a logging system. We can set restrictions for access to the database in order to maintain the confidentiality of the data. They are also known as SQL database. Some examples of relational databases include MySQL, which is organized into physical files optimized for speed. ORACLE, which is based on the principle of relational DBMS. SQL query are required to access, modify and get desired data from the database.

Advantages:-

* Speed of Relational database is high because of its ease and simplicity.
* Database can be protected by using credential.
* Data redundancy, inconsistency, access control can be gained.
* Data independency.
* Reduce application development time.

Disadvantages:-

* Storing a large number of data required huge amount of storage.
* Need to have knowledge of SQL(Structured Query Language).

**Question 1b:**

Apart from qualification which was also a great shock to me, the technical knowledge required in the field SQL, ETL, data modelling, and at least one programming language (e.g., Python, C++, C#, Scala, etc.), Understanding of business intelligence and data analysis tools (i.e. PowerBI, Tableau, etc.) , Strong proficiency in Excel/Google Sheets and financial modelling, T-SQL, Visual Basic, Python, MongoDB, SQL Server BI Suite (SSIS, SSAS, SSRS). This all where the teams that I was came across during my search.

* Able to perform advanced SQL queries on very large datasets and to gain proper insights from that.
* Create and maintain an analytics dashboard using various data visualization tools like R studio, Power BI or Tableau.
* Need to gain knowledge in field of ETL/data wrangling tools.
* Should have good hands-on different types of databases like MySQL, MongoDB.
* Reporting tools like Excel is one of the requirement skills, as most of the data which is extracted from external sources are in .csv format.

**Question 2a :-**

Imagine that we are building a house in a small piece of land, we need some planning, to design a house based on our need. So the architecture design a blueprint of the house based on the requirements of the land owner. For this architecture needs to design a model for better understand of how the house would look like. It is very essential to design any model to build any house. Similarly it is very important to design a Data model to build any database for a particular enterprise. Data model is a visual representation of an enterprise’s data and connection between them. A Data model represents a framework of what the relationships are within the database. This framework provide the structure which will support the analytical needs of the decision maker. Building a data model is an important step in the design of the data warehouse. Data model can be design into three stages which are-

Conceptual data model - Its is also called as domain model. These are a square shape that are connected by a line. The square shape represents an entity and the lines represents a relationship. These model is mostly used in the beginning of a new project. This model is highly abstract that means it does not contain any detailed information.

Logical data model - Logical data model is a detailed version of Conceptual data model. Additional details such as attributes for each entity, Key attributes are added in the data model. Below figure is an example of Logical Data model.

Physical data model - Physical data model looks almost like a Logical data model, however there are some significant changes. In physical data model entity is referred as tables and attributes are referred as columns. The columns names are database compatible names. Besides attribute, data types are also mention. Below figure is an example of Logical Data model.

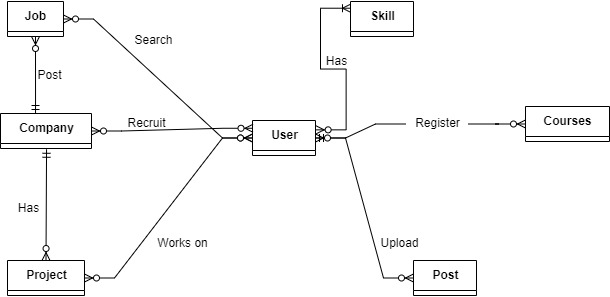
I have made this ER diagram using draw.io where I have mention various relationship between entities.

**LinkedIn Case Study:**

The social media like LinkedIn is a tool which help the user to search for jobs and the recruiters to collect the applications from them and the company also can post the available position in their area and the user can see those position and can also, apply through LinkedIn also. The user can create their own profile to attract the recruiters. The users can also develop new skills and enrol in certification courses to make their profile strong so can have higher preference with respect to others. The users in the LinkedIn are of two types one is a normal type and the other is the premium type in which the user can see who has viewed the users profile and has some benefits over the normal user as they have to buy subscription plan for the month or the year depending upon the user availability of the user.

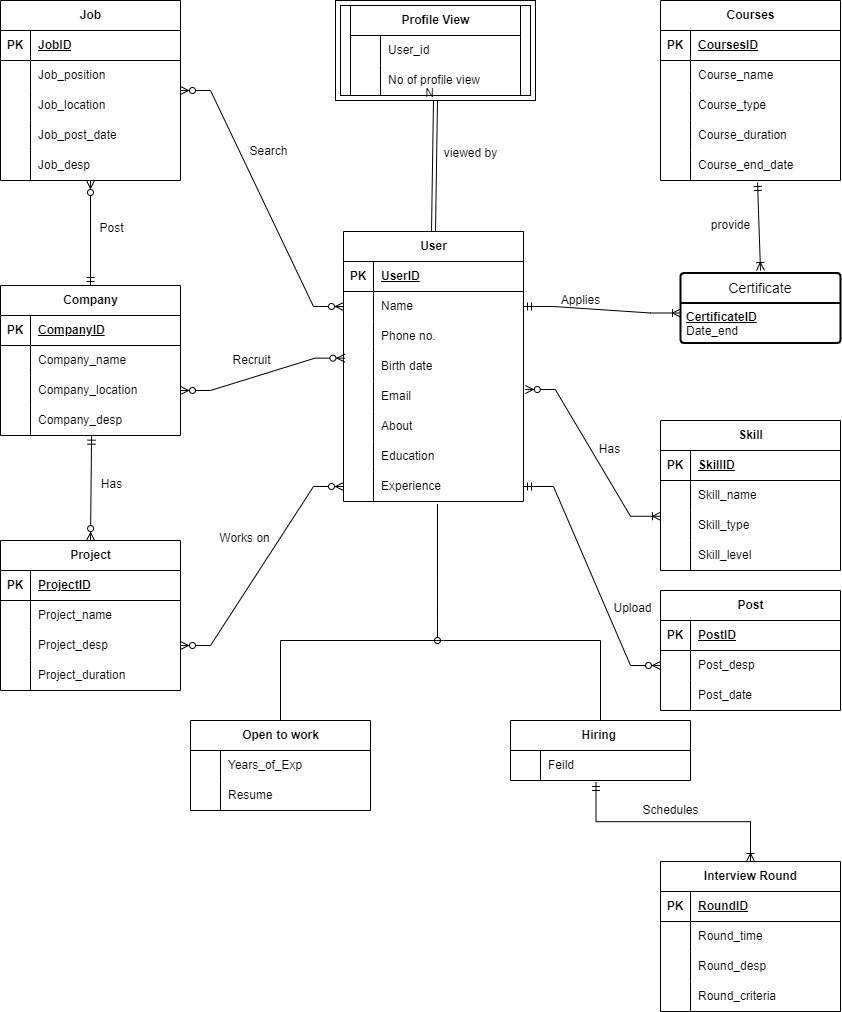
1. Enterprise data model:-

It is a high level entity and relationship for the organization but it over view of the project data model just for upper level and easy to understand for the people from other department.



**Fig 1:- Enterprise model for LinkedIn**

1. Project data model:-



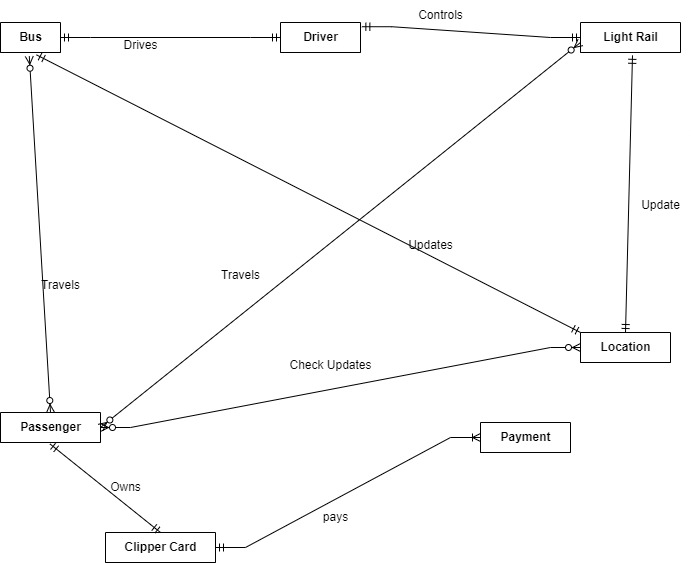
**Fig 2 :- E-R Diagram for LinkedIn**

**Question 2b:-**

**Valley Transport Association (VTA) Case Study:-**

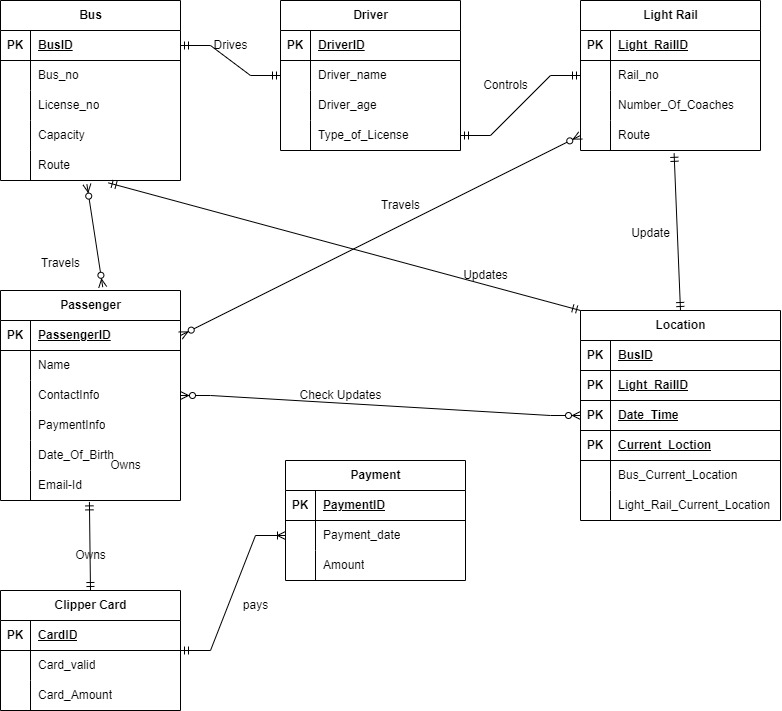
The VTA is a public transport which severs the people with their transports and to make their life easier and simpler. So, the system offer buses service and light rail service for the transport. The government also provide as system of Clipper Card with is like a recharge card system in which the user just have to load the amount of money they are going to use in future and the fee for each trip is fix redragging the distance the user will be travelling.

1. Enterprise data models:-



**Fig 3:- Enterprise data model for VTA.**

1. Project data models:-



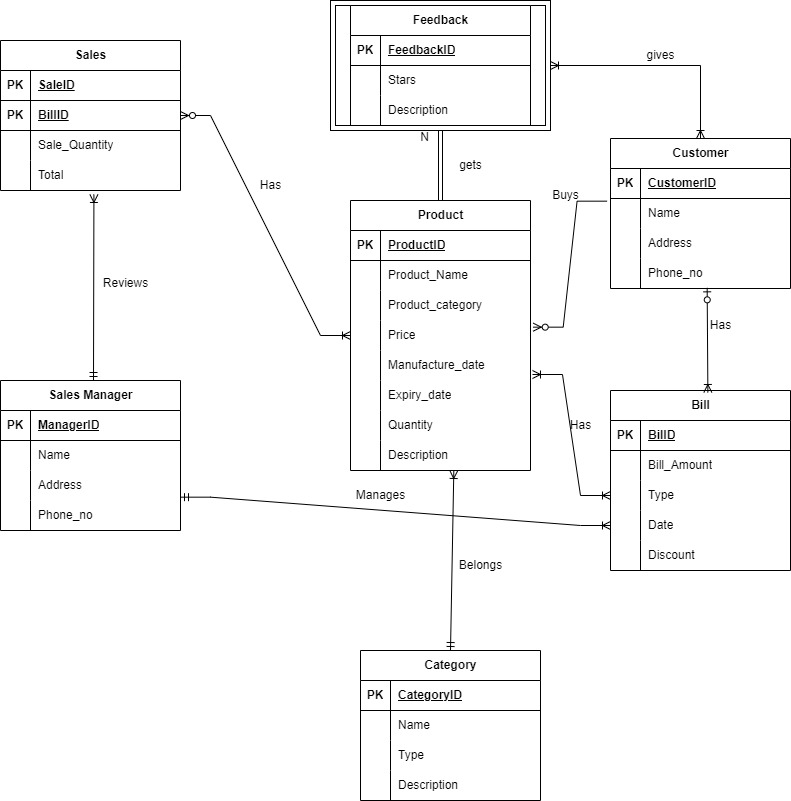
**Fig 4:- E-R Diagram for VTA**

**Question 3:-**

The first the scenario from the real world is the Retail store and the idea of studying it came from my personal experience from the Retail store in front of my apartment called “Whole food”. So, in the retail store the its has customers, managers and the products to manage by the database and the feedback from the customer helps the store to improve it service to the customers, this will also help to gain customer satisfaction. The second part is the billing part which will be under the surveillance of the sales managers.

* **E-R diagram Retail store:-**

Below is the ER diagram of retail store. We have multiple entities like Sales, Product, Customer, Store Manager, Bill, Category. Each entity is has some relationship to other entity. We can see that there are multiple cardinality. Every store has multiple customers and customer buys one or more than one product. Every store has a data model to keep a track of sales of the products.

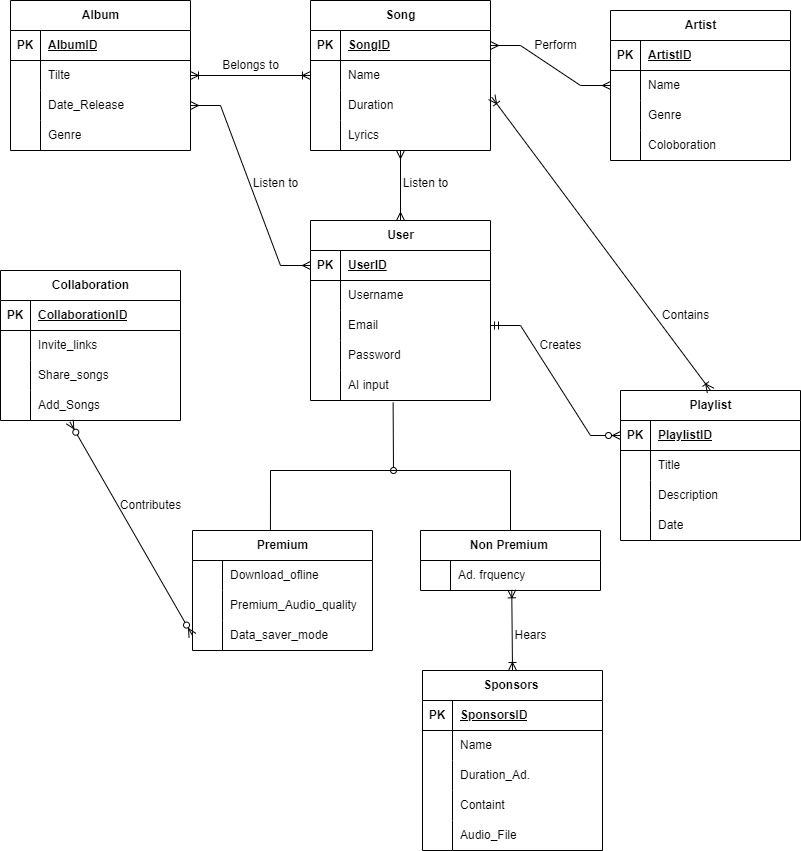


**Fig 5:- E-R diagram for Retail Store.**

In the second scenario, I have decided to do a study in the Spotify application which s basically a database which has a good storage of songs apart from this the main reason to do it because I use it every day and nearly 8-9 hr per day. So, it has a songs database from which the end user can hear it apart from this user can make his own playlist as much as he wants or just can search form the bar to listen them and the algorithm will automatically make the queue according to your inputs.

* **Spotify:-**

Below diagram is the E-R diagram representation of the Spotify application which has Songs, Artist, Album, User, Collaboration, Playlist and the sponsors. So, the model they use is also a subscription base so we have two type of user like premium and non-premium user, both types have different access to databases and features to the application.

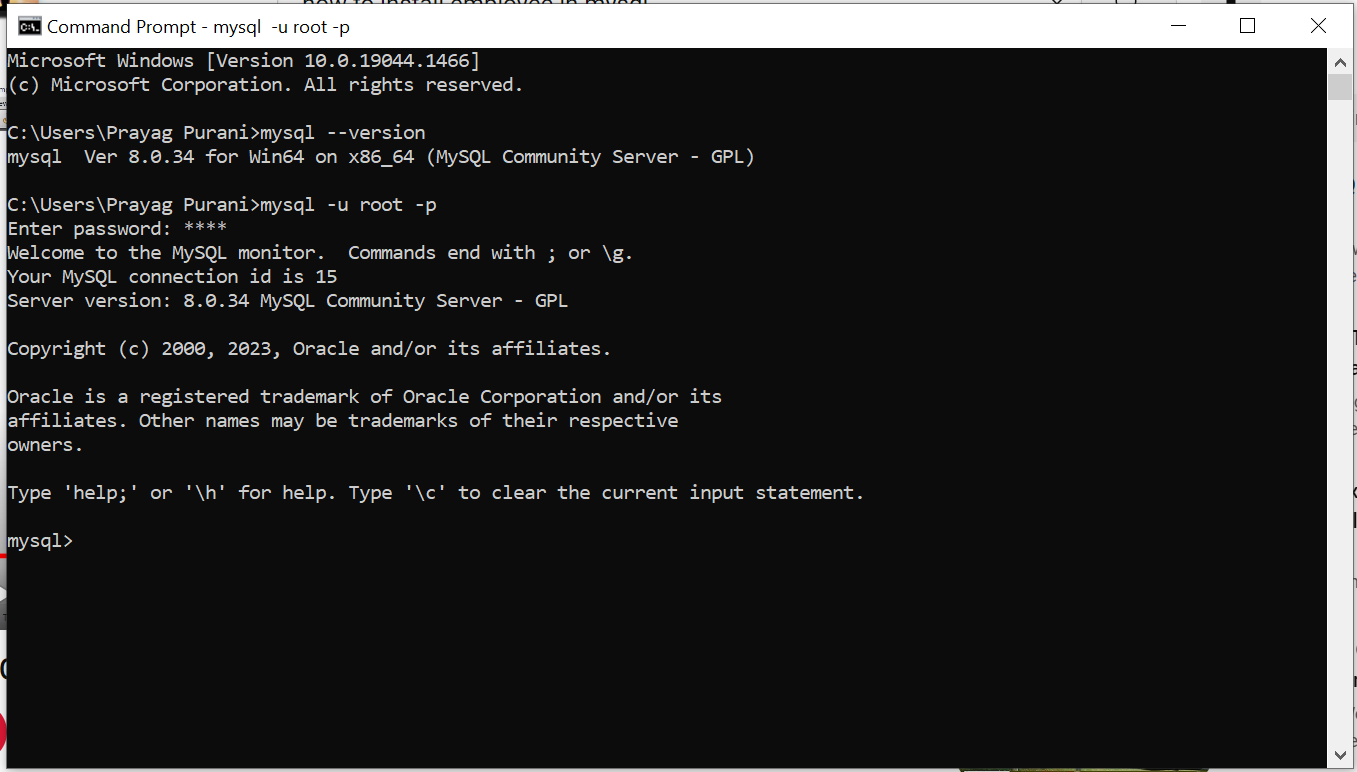


**Fig 6:- E-R diagram for Spotiy.**

**Question 4 :-**

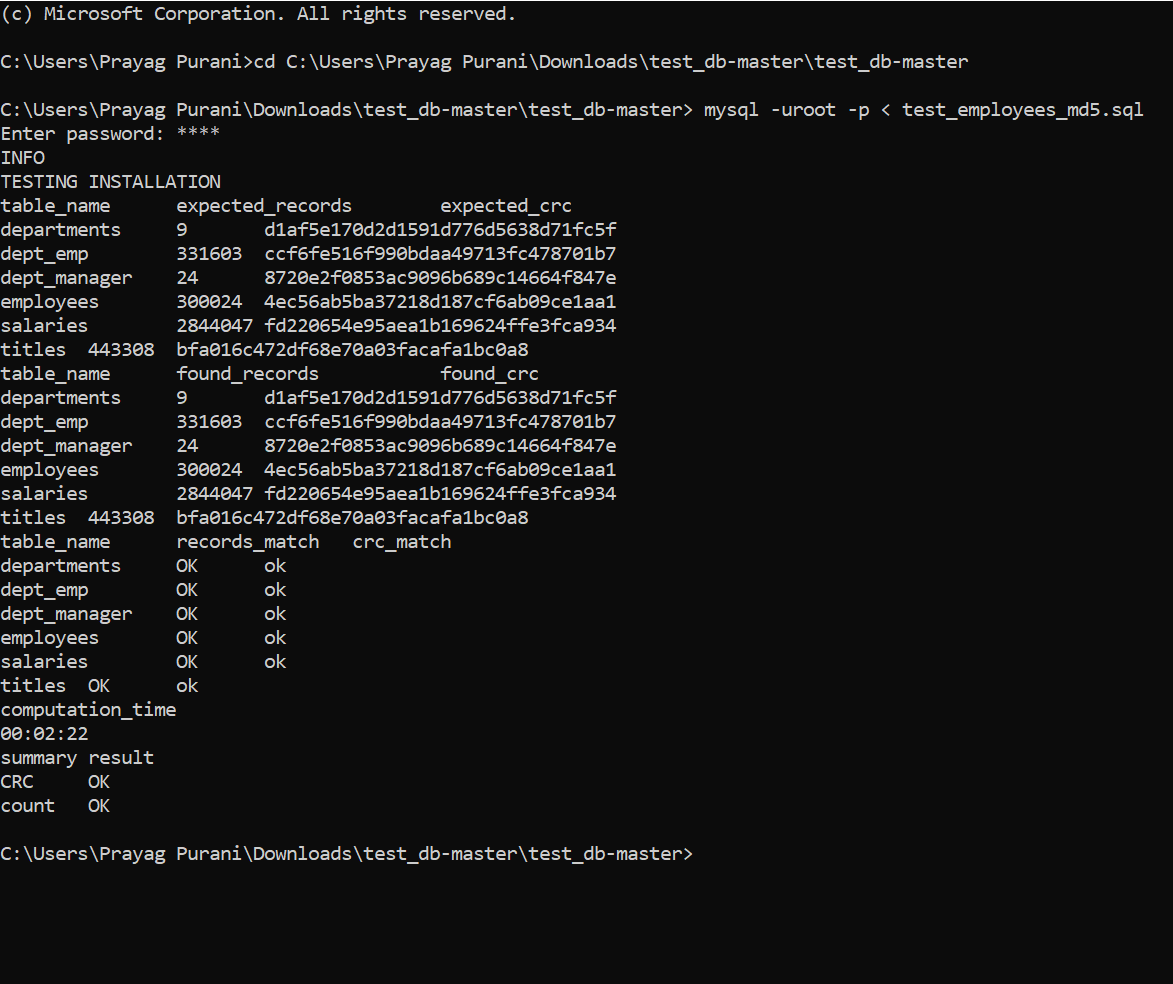
I have successfully installed the MySQL on windows operating system, by follow the link and refereeing to some videos online.

Starting SQL:-



**Fig 7:- MySQL installed**

Know adding the employee database to root user:-



**Fig 7 :- Importing employees database and testing it**

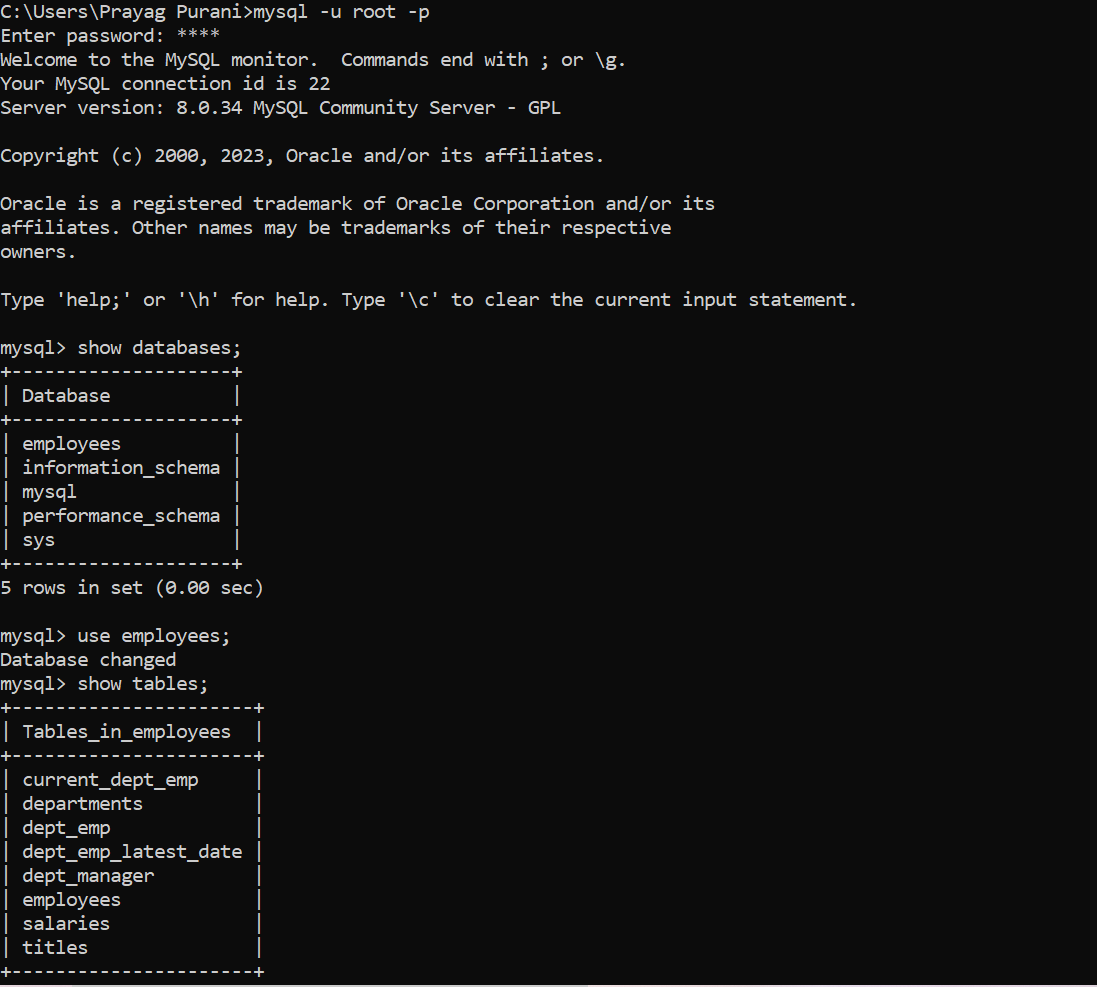
Executing the queries:-

Show databases;

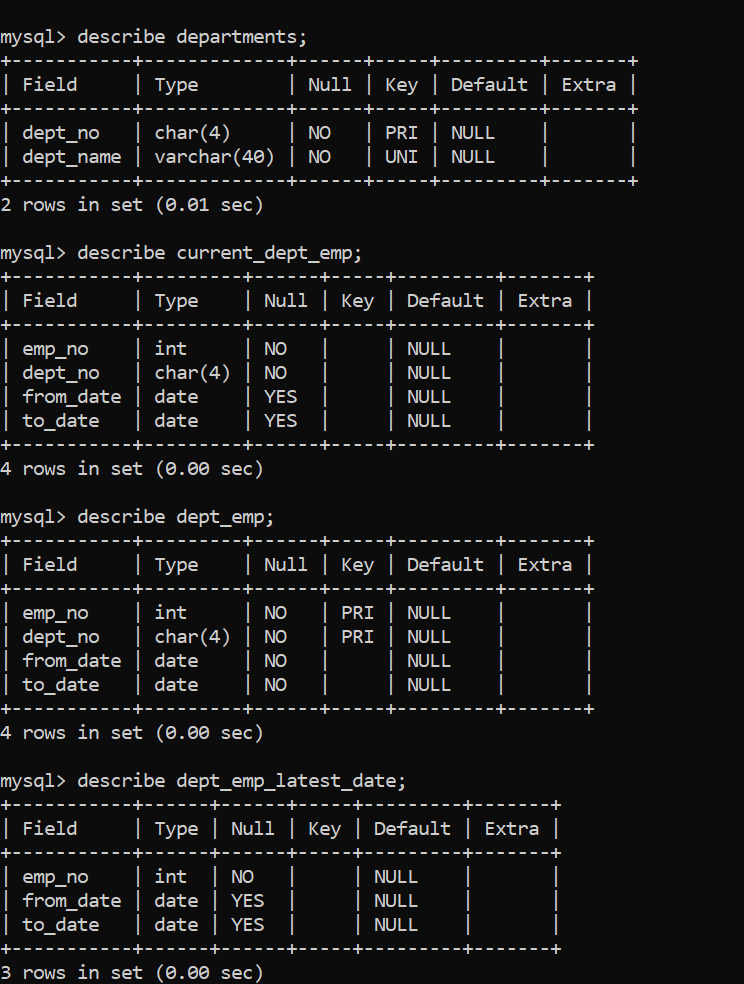
Use employees

Show tables;

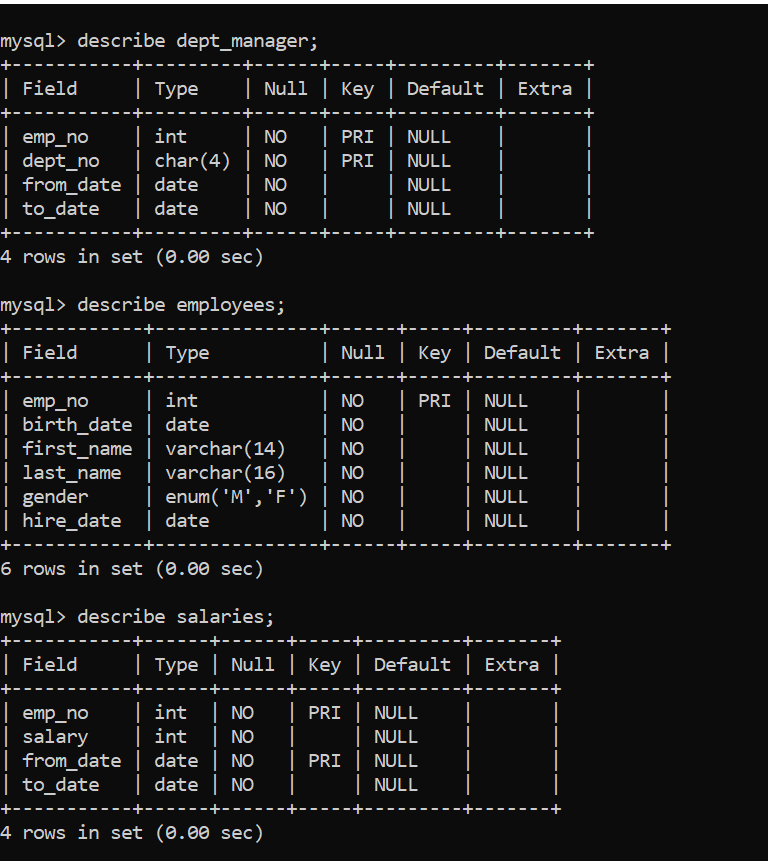
Describe departments; and so on



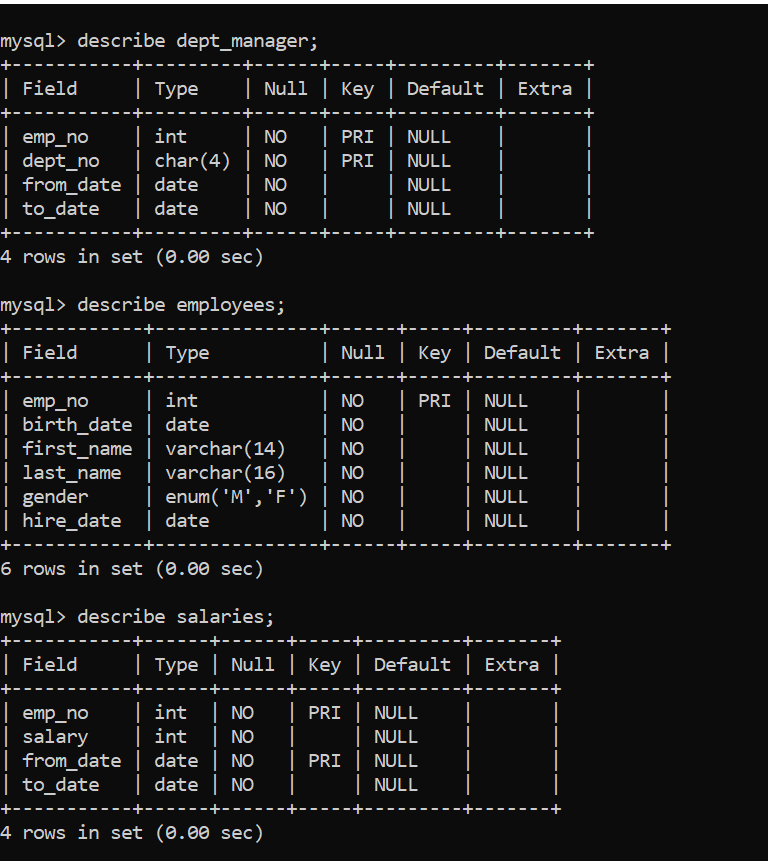
**Fig 8:- Query execution**

****

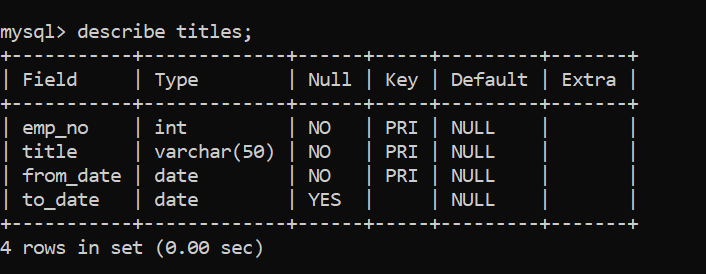
**Fig 9 :- Query execution**



**Fig 10 :- Query execution**

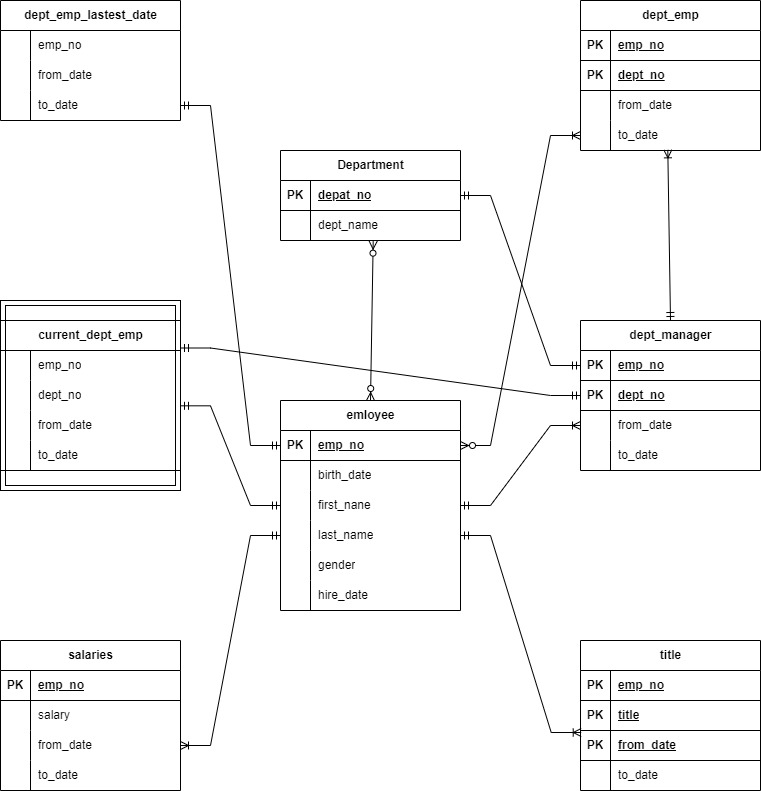


**Fig 11 :- Query execution**



**Fig 12:- Query execution**

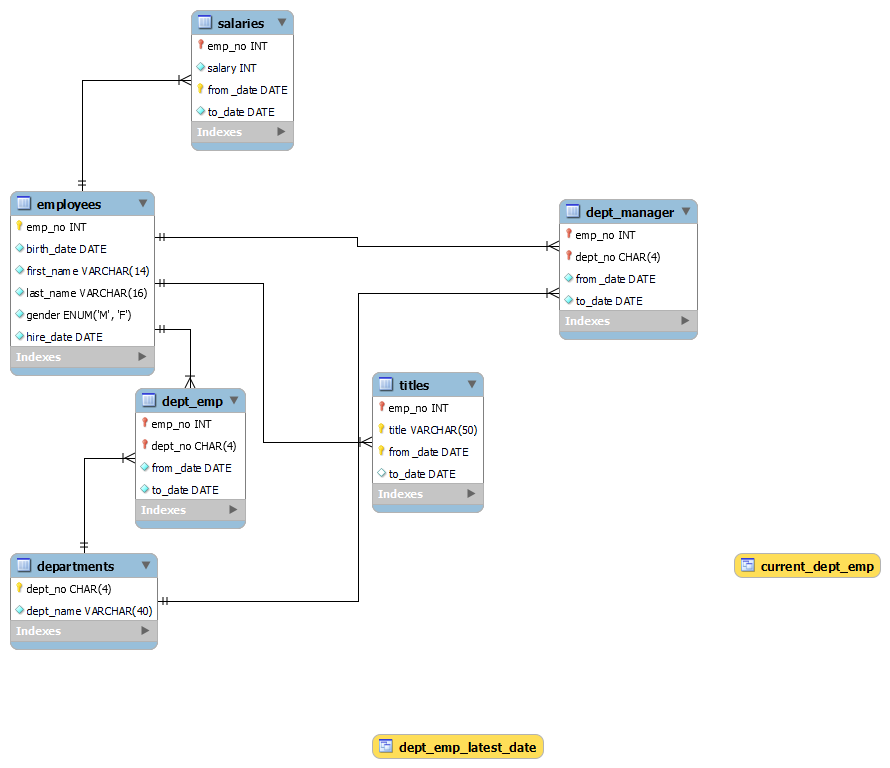
**ER diagram:-**

****

**Fig 13:- ER diagram of employees database**

**Question 5a:-**

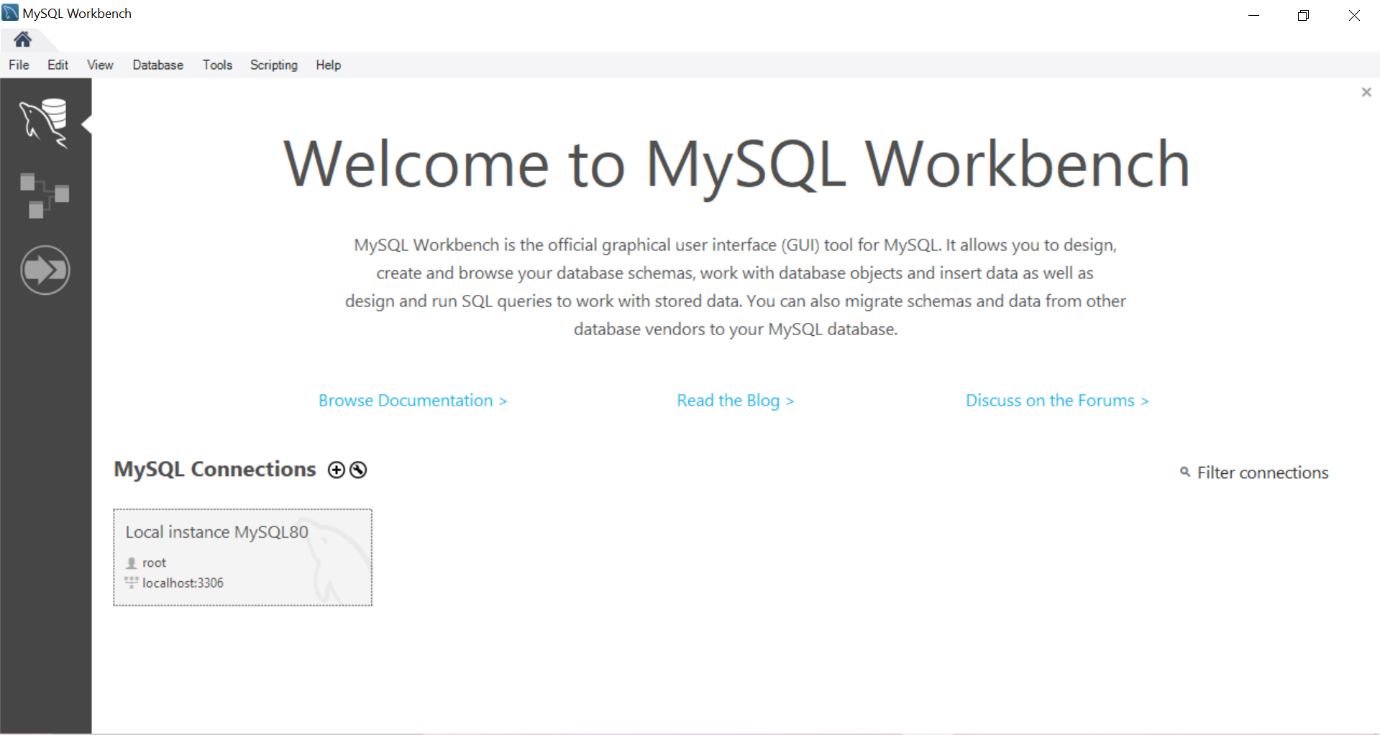
I have successfully install MYSQL Workbench for windows from the link provided in the task, also followed the steps to generate ER diagram. Below is the result for the same.

****

**Fig 14:- ER diagram of employees database Reverse Engineering**

**Question 5b:-**

MySQL Workbench is an Integrated Development Environment (IDE) for MySQL .It is an application which allows us to run MySQL queries. It is an open source application which has various option available to work with the database. The current version of MySQL workbench is ‘MySQL Workbench 8.0.33’. Below are some of the features of MySQL Workbench.



**Fig 15:- Workbench Home page.**

* MYSQL Workbench functionality covers:
  + SQL Development
  + Data Design
  + Server administration
  + Data Migration
  + MYSQL Enterprise Support
* The storage function is used to configure a custom path to the table storage and data file.
* The performance dashboard is used to check if the SQL server is experiencing poor performance, longer query times or storagerelated issues.
* The performance schema report is a tool that allows you to view and analyze the data collected by the performance schema.
* MYSQL workbench provides DB Doc which can be used by a developer to give point-and-click database documentation.

**References:**

<https://blog.packagecloud.io/ten-things-about-binaries-you-need-to-know/#:~:text=A%20binary%20file%20is%20a,for%20the%20file's%20text%20content>.

<https://www.techtarget.com/whatis/definition/binary-file>

<https://www.techopedia.com/definition/9707/text-file#:~:text=A%20text%20file%20is%20a,flat%20files%20or%20ASCII%20files>.

<https://en.wikipedia.org/wiki/Text_file>

<https://blog.hubspot.com/marketing/how-to-use-excel-tips#:~:text=Excel%20is%20used%20to%20store,%2C%20budgets%2C%20or%20editorial%20calendars>.

<https://www.microsoft.com/en-us/microsoft-365/excel>

<https://docs.oracle.com/cd/E19078-01/mysql/mysql-workbench/wb-intro.html>

https://dev.mysql.com/doc/workbench/en/wb-intro.html

<https://www.youtube.com/watch?v=-ksz8J6FQ0A>

<https://www.youtube.com/watch?v=qBBXBZXKi0w>

<https://www.youtube.com/watch?v=fUK94jOFwBc>