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

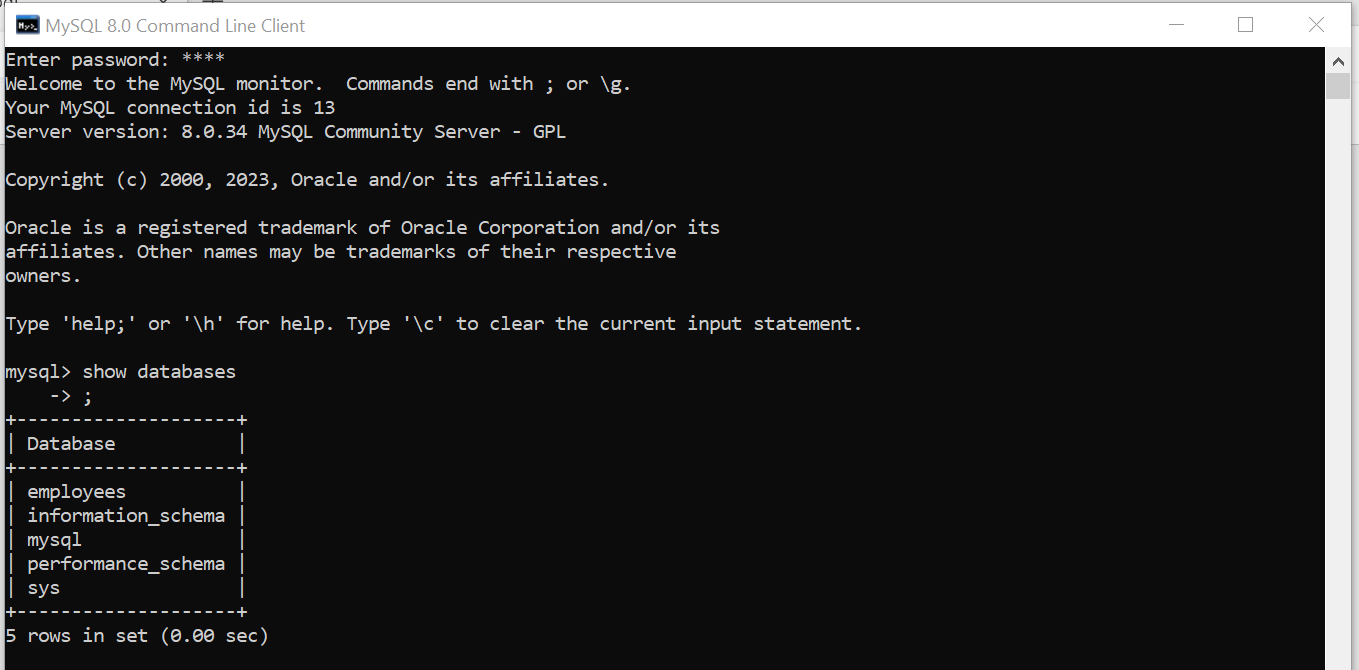
**Homework - - 2**

**Name :- Prayag Nikul Purani**

**SJSU Id :- 017416737**

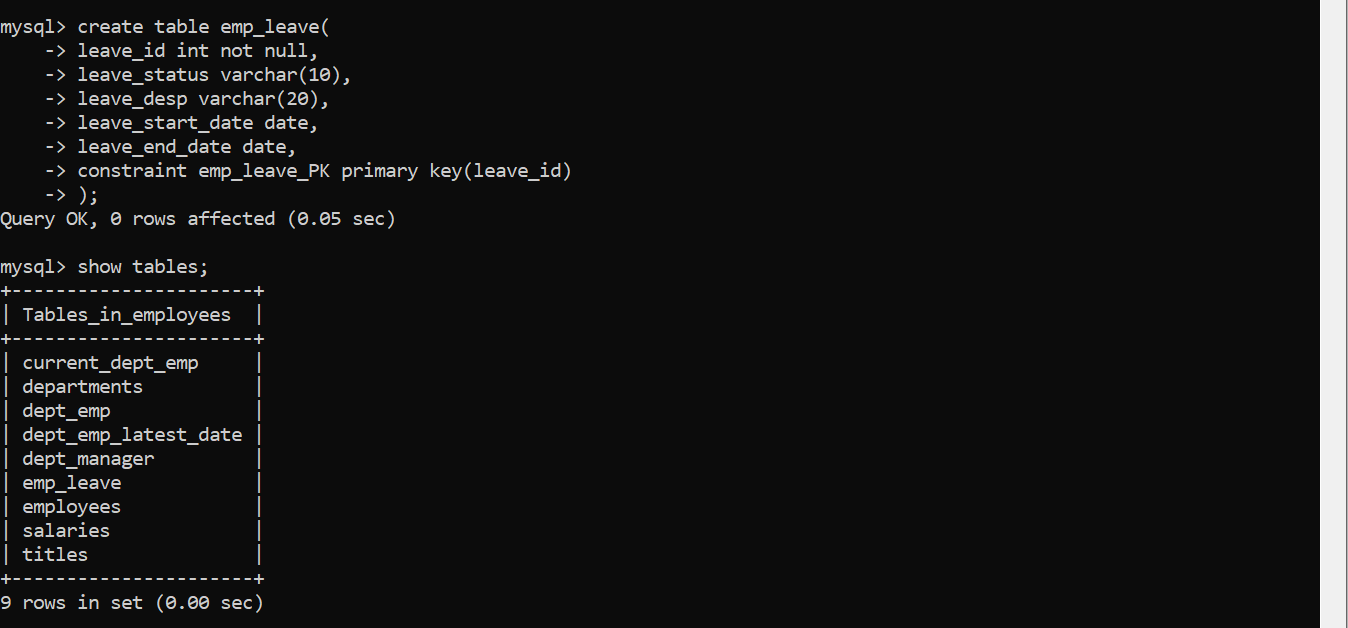
**Question 1 a:**

**Write one SQL DDL statement each for any 5 of the tables in the employees database that you imported into MySQL in HW1.**



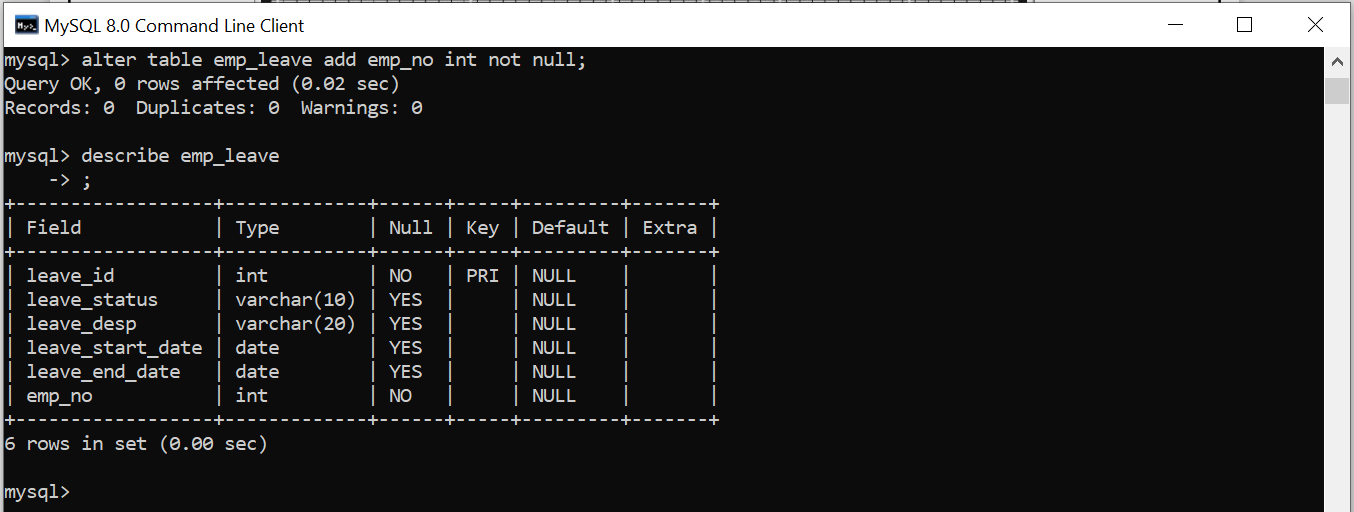


**Query 1:** Creating a new table emp\_leave for the leave status of the employeess as DDL query for already creadted tables will cause an error in MySQL. **CREATE TABLE COMMAND.**

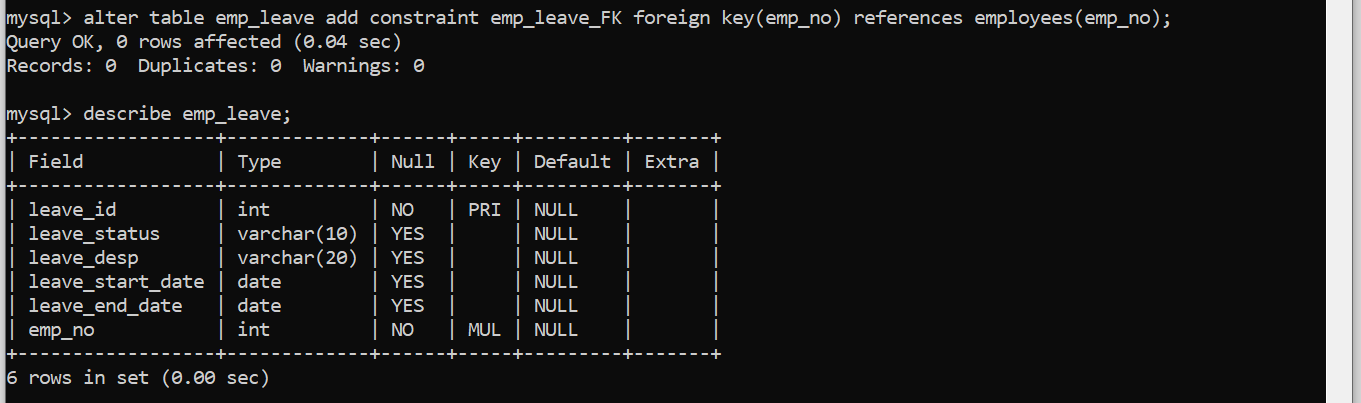


**Query 2:** After creating a new table emp\_leave for the leave status of the employeess we have not added a foreign key which will be employee no from the employee.

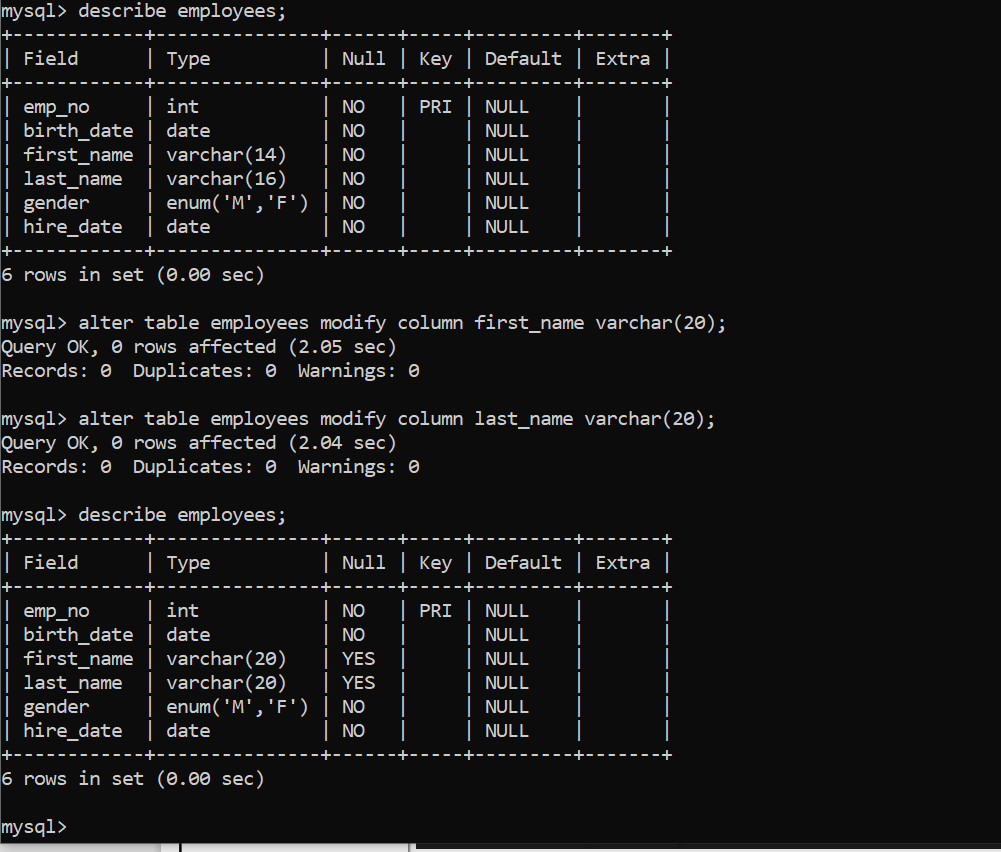
So for this process we need to first add emp\_non attribute in the emp\_leave table. **ALTER WITH ADD COMMAND.**

****

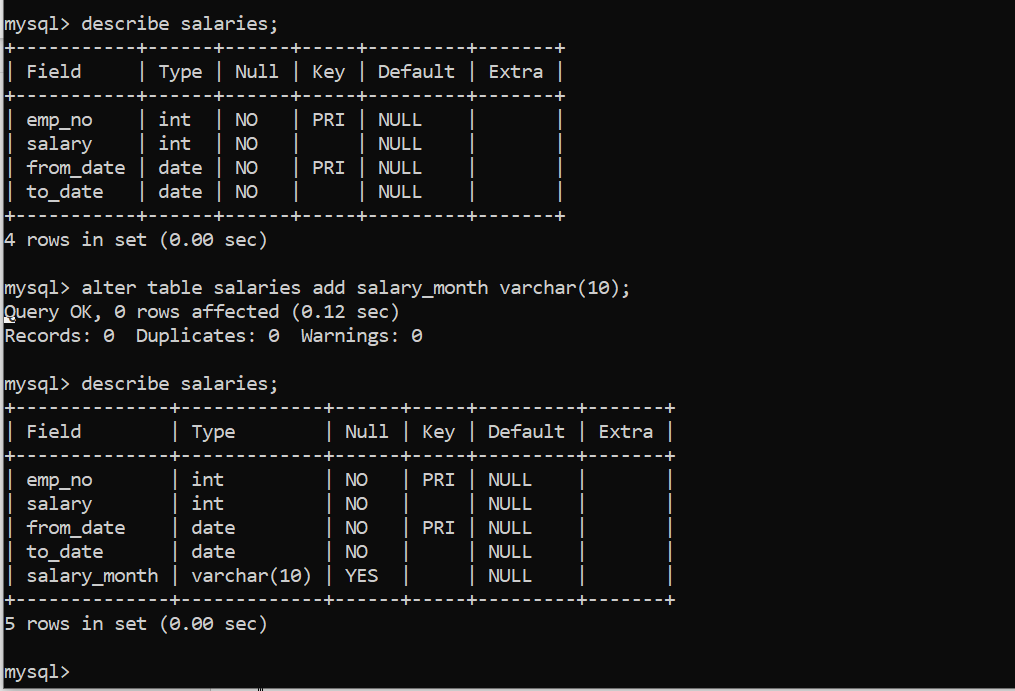
So, the second step is to add constraint of foreign to the above inserted column. **ALTER + ADD CONSTRAINT COMMAND.**



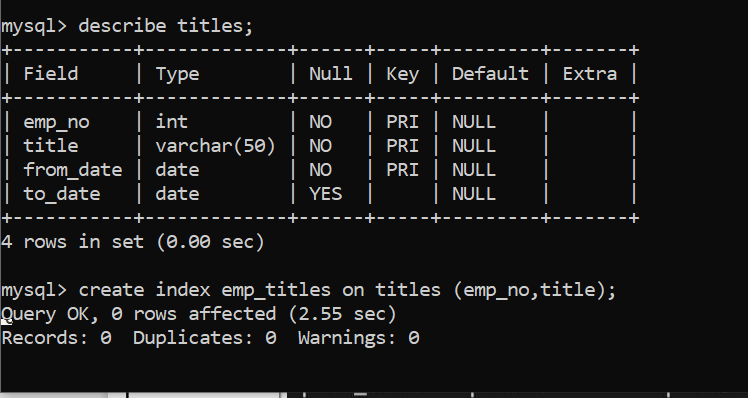
**Query 3:** After describe command on employee we can see the length of the first\_name and las\_name is 16 and 14 respectivily so, modifying it to length 20. **ALTER + MODIFY COMMAND.**



**Query 4:** Adding a new column to salaries table naming it as salary\_month with length 10. **ALTER + ADD COMMAND.**

****

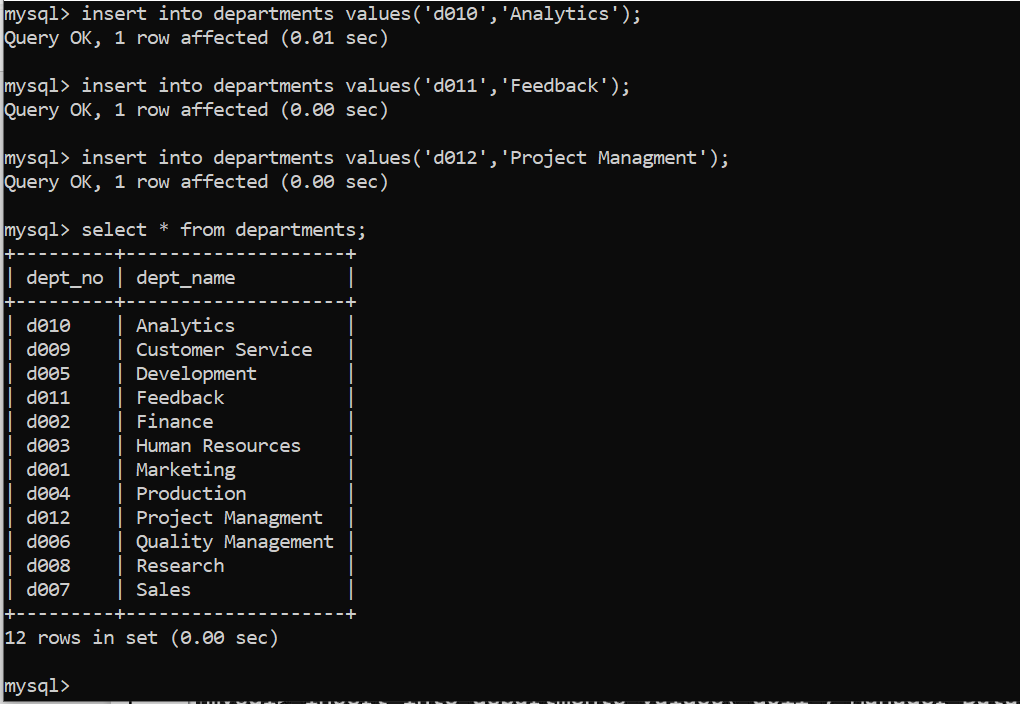
**Query 5:** The SQL statement you've provided is creating an index called "emp\_titles" on a table named "titles" with two columns: "emp\_no" and "title". **CREATE + INDEX COMMAND.**



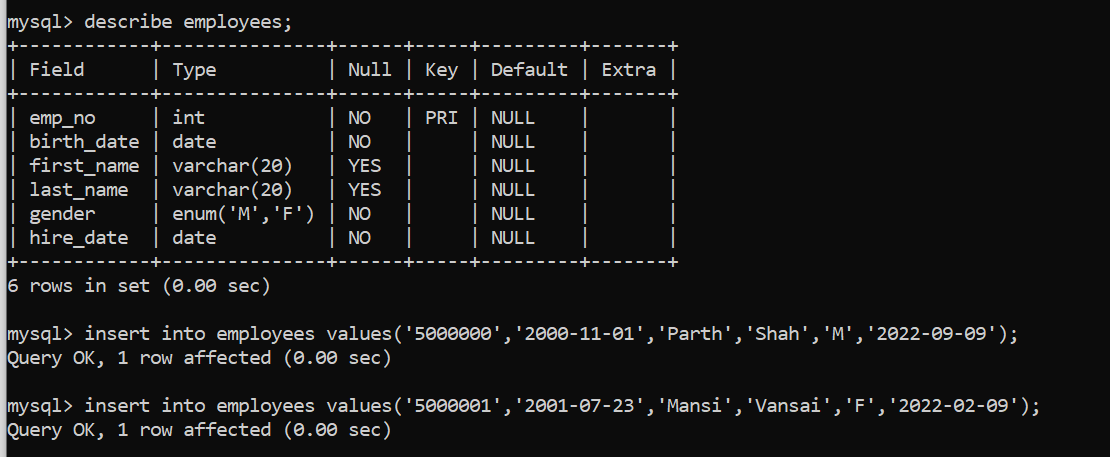
**Question 1 b:**

**Write one SQL DML statement each to populate each of the above 5 tables with one or more rows of data.**

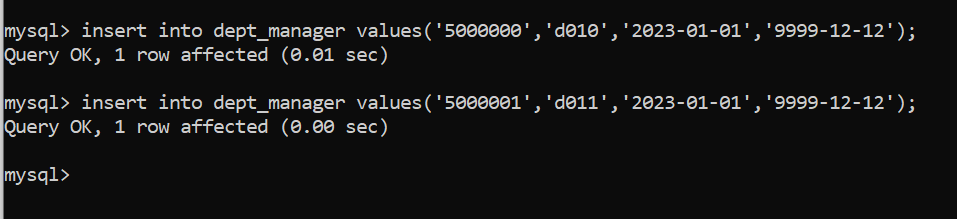
**Query 1:** Inserting into department tables

****

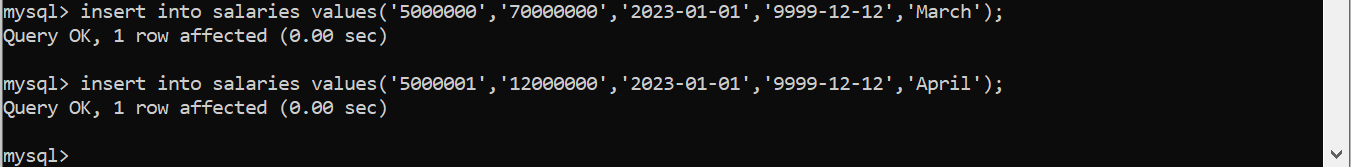
**Query 2:** Inserting into employees tables.

****

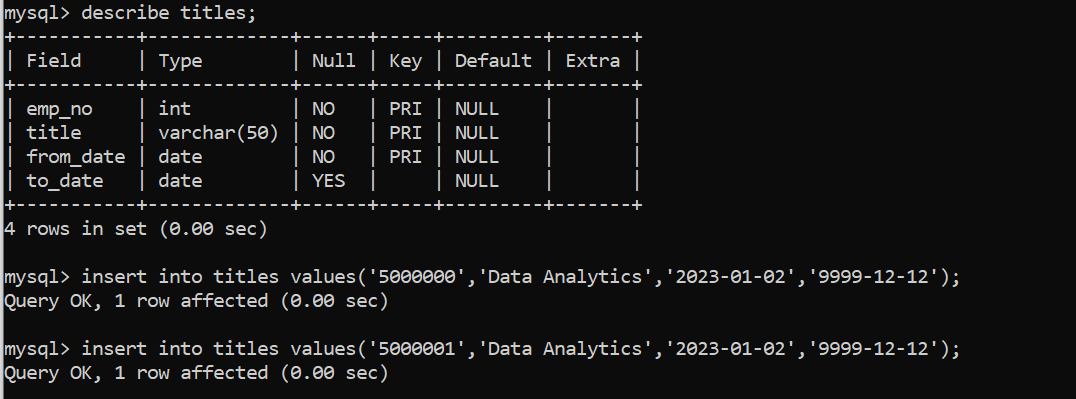
**Query 3:** Inserting into dept\_manager values.

****

**Query 4:** Inserting the values into salaries table.

****

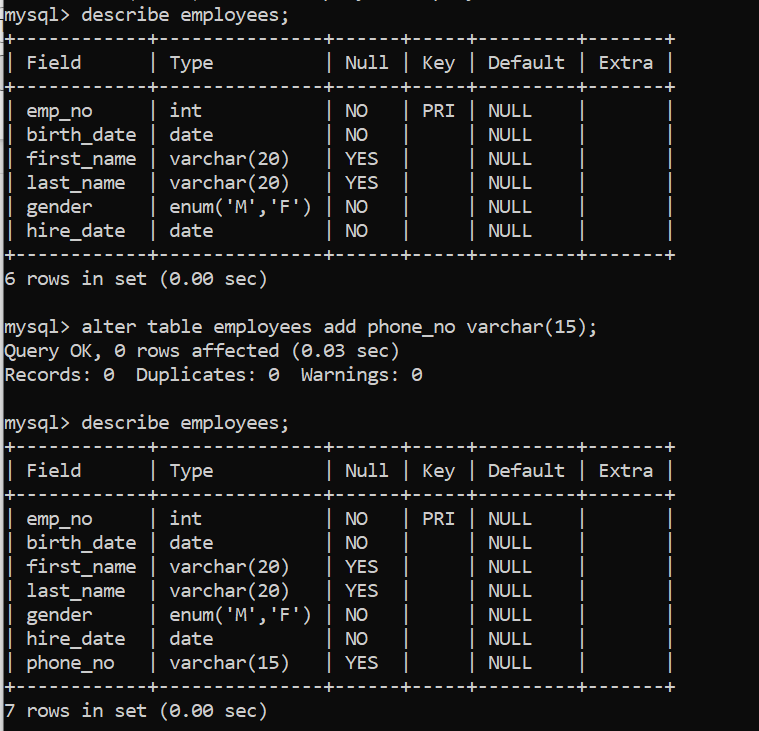
**Query 5:** Inserting into titles tables.

****

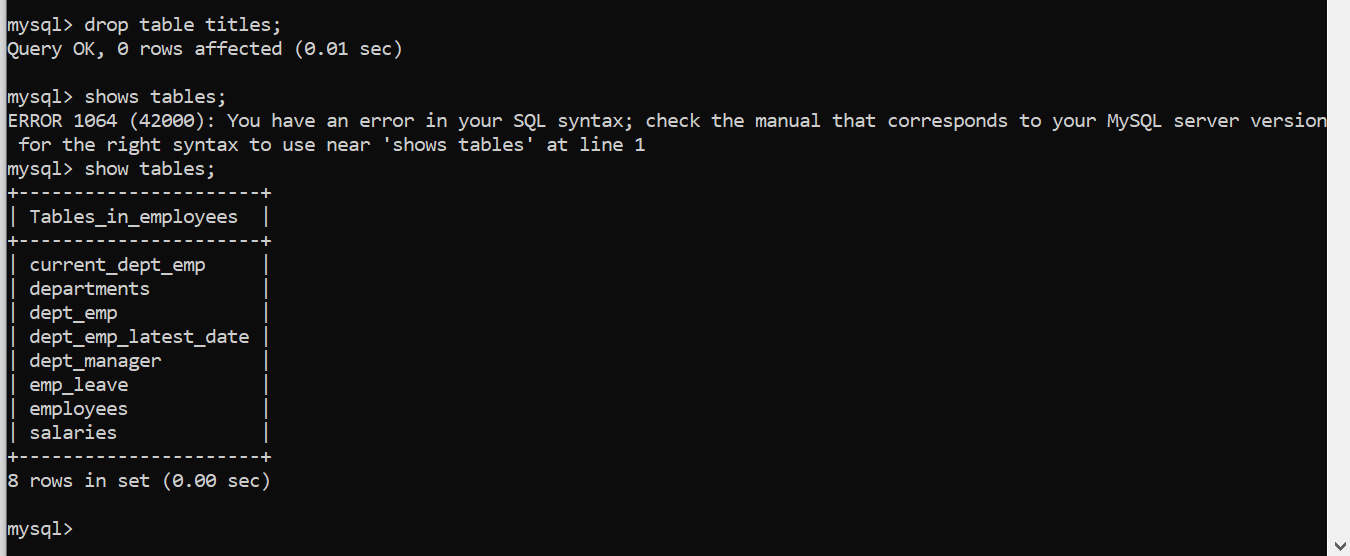
**Question 1 c:**

**Write SQL statements for each of the following scenarios as applied to the employees database:**

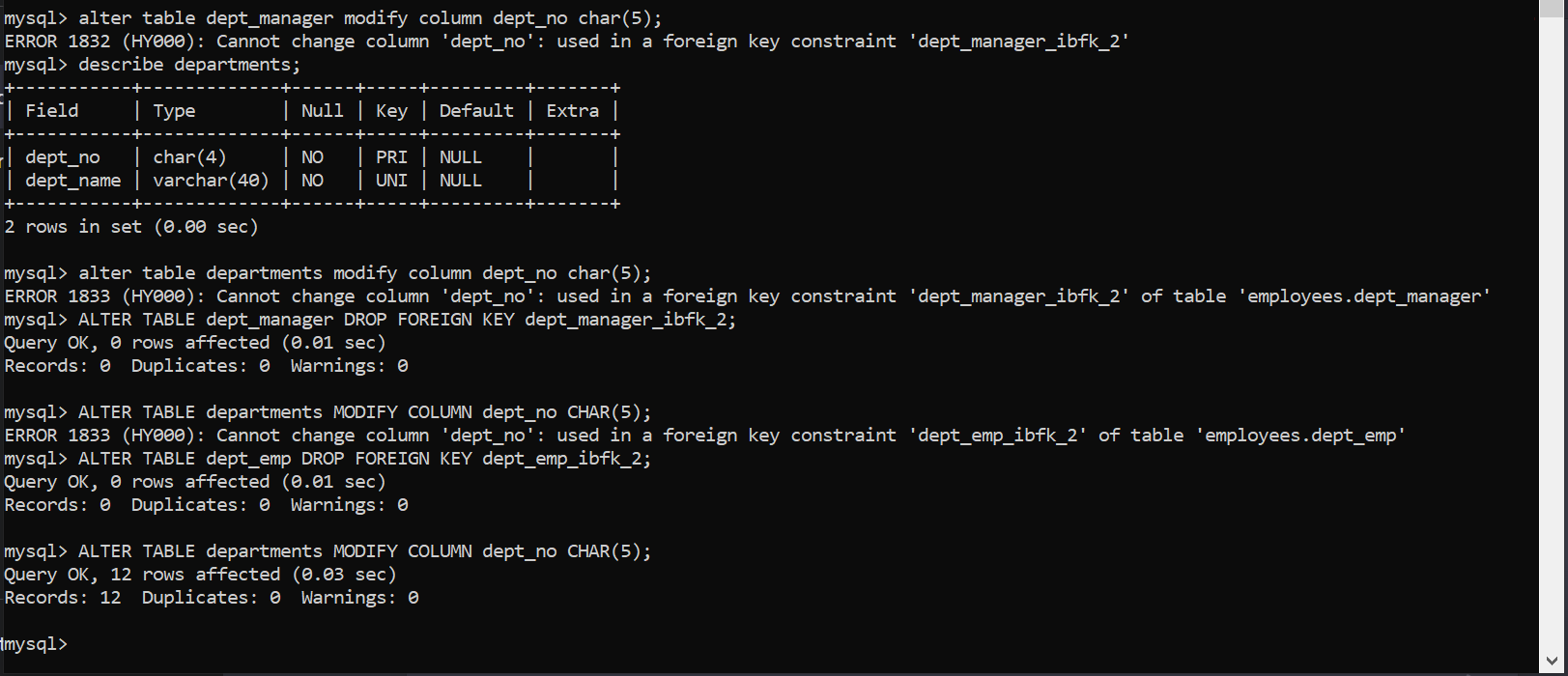
1. **Add an attribute, Phone, to the employees table.**

****

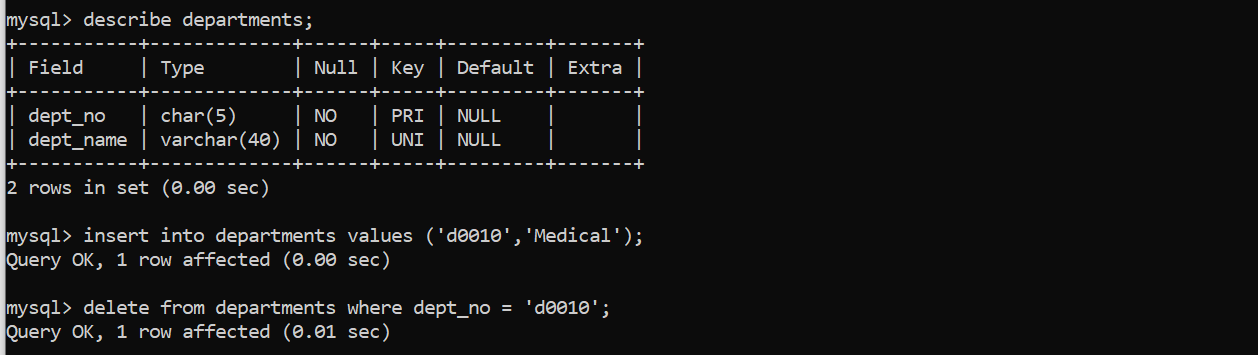
1. **Remove the titles table**

****

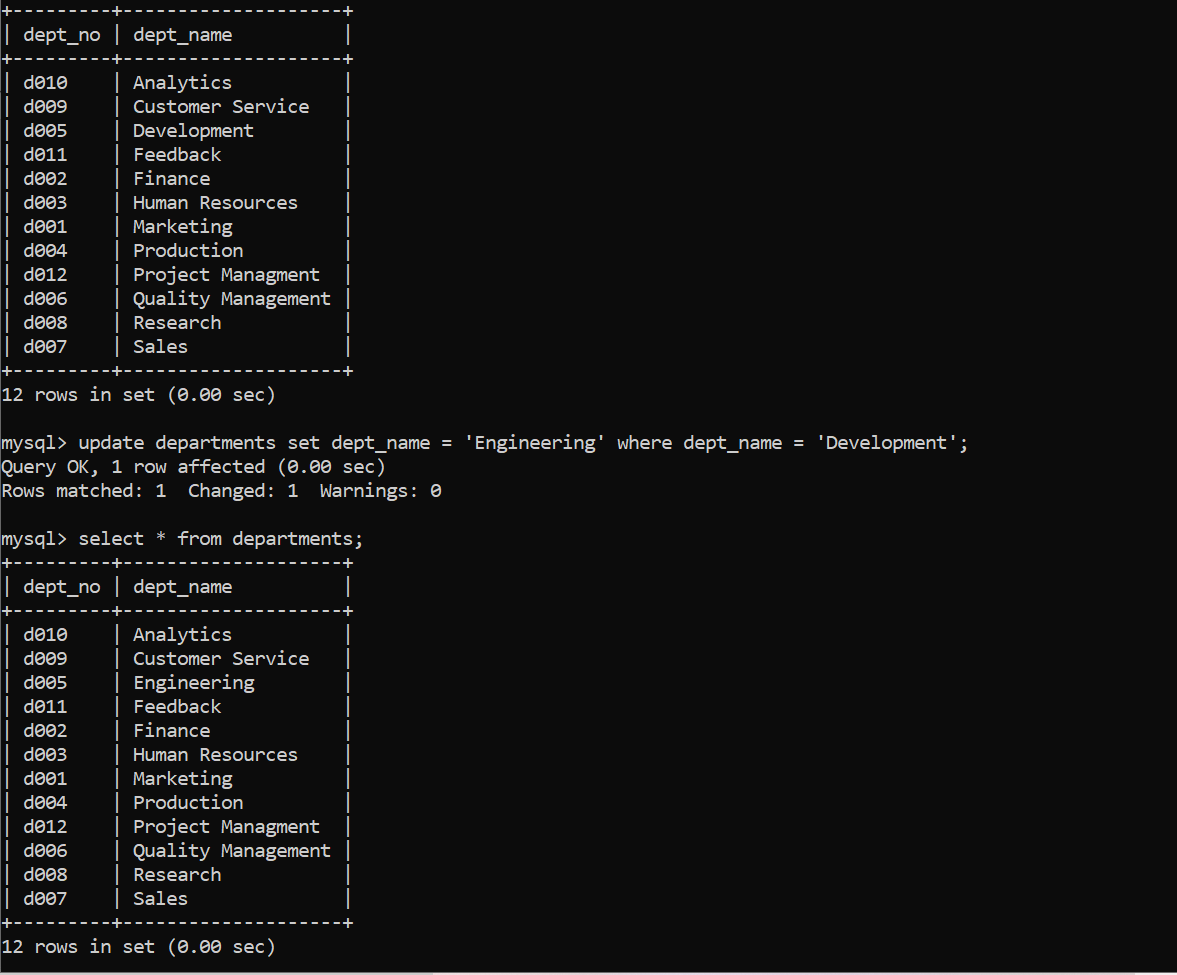
1. **Change the dept\_no field from 4 characters to 5 characters.**

****

1. **Add a new department called “Medical” with number d0010 and later delete it.**

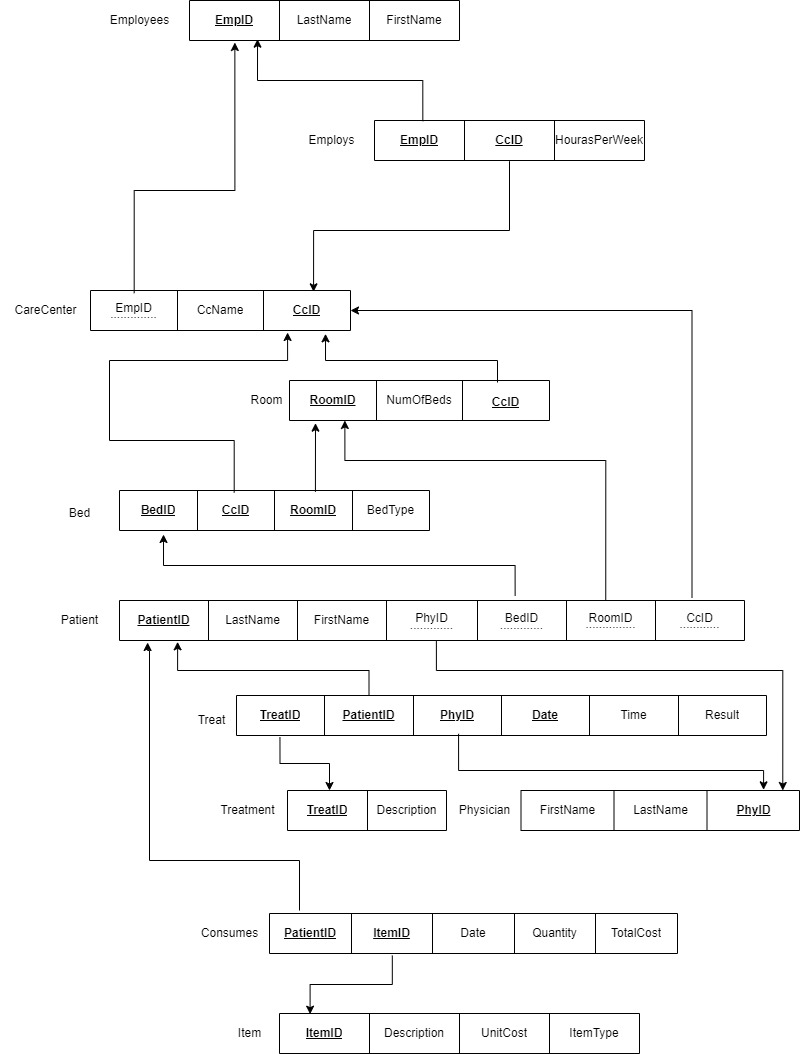
****

1. **Modify the department name from “Development” to “Engineering”.**

****

**Question 2 a:**

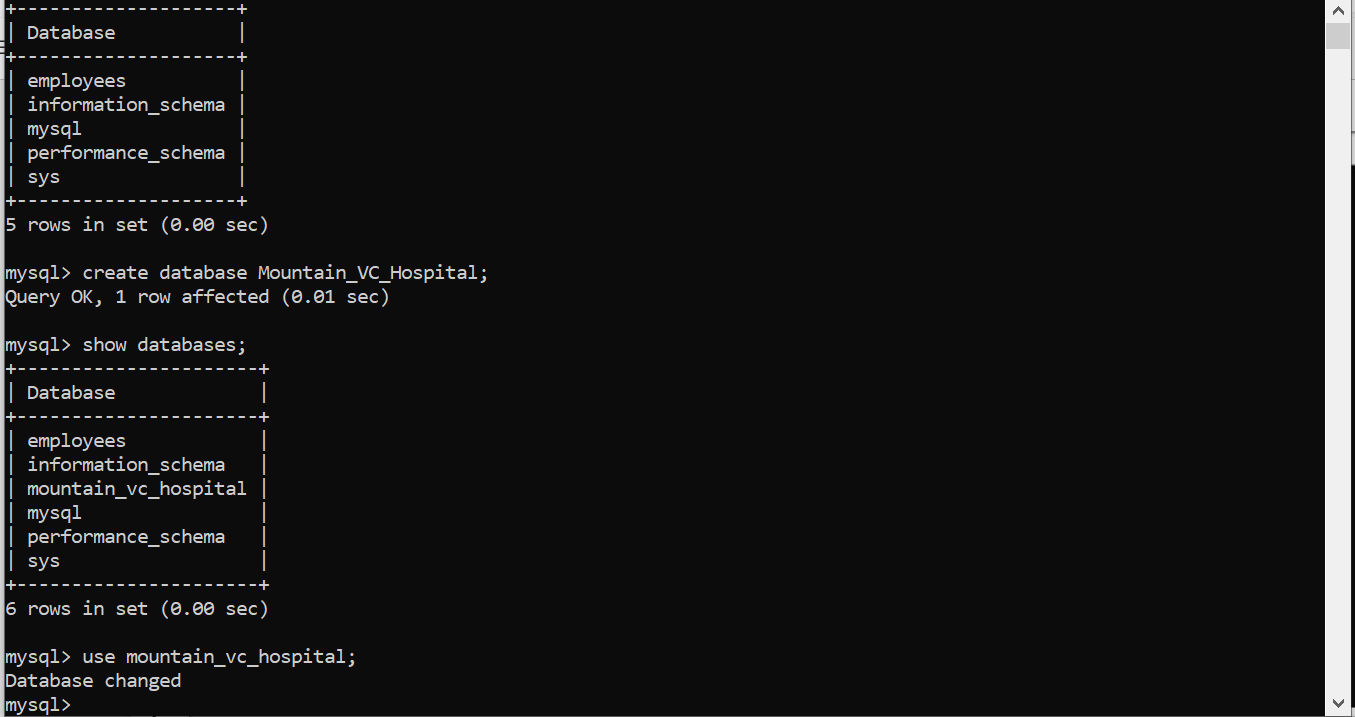
**Convert an ER diagram for a hospital, into a relational schema. The MDM textbook describes the case study – search for “Mountain View Community Hospital” in the book for more details. Please feel free to make any corrections or changes that you may deem necessary to the above ERD before generating the relational schema.**

****

**Question 2 b:**

**For the above hospital relational model, write SQL DDL statements and assertions necessary to create the tables and ensure referential integrity and other constraints.**

Now it’s required to create a new database called “Mountain View Community Hospital”



Now the creation of tables with the referential integrity and other constraints.

First table to create is Employee after that Care Center will be created and then their intersection table Employs will be created.

**Question 3b:-**

The highest normal from discuss in class was 3rd normal form, there are many more from beyond 3rd NF but going for those is not efficient and will reduce the time as the number of tables increases so it will lead to increase in number of hops so will reduce efficiency, so its advice to leave some redundancy in the databases.

So, we can tell Hospital and Employee databases are in 3rd normal from.

1. In both the databases there is no multivalued attributed and it satisfy all the condition to be a relation so it’s in a 1st NF
2. In all the tables in each database there is no partial dependences so, each attribute is fully depended in the primary key so we can say that it’s in 2nd normal form.
3. Now to check for 3rd NF we can say that it has no transitive dependency so the final step is also satisfied to it’s proved that both the databases are in 3rd normal form.

**Question 5 a:**

We have interviewed Mr. Calvin Dani. He has great experiences in the industry as he had more than 2 years of experience in India and he is working as RA in Database and Information at Science Lab at Santa Clara University.

<https://www.linkedin.com/in/calvin-dani/>

1. Can you please introduce yourself?

Hi, my name is Calvin Danny. I'm a master's student in Santa Claus University. I'm currently in re at the Database and Information Science lab at Santa Claudia University and working in projects in A TV and couch spaces.

1. What is/was your roles and responsibilities in your company?

The main role as a research assistant is to find new ways to implement to find a solution to our problem or one of such problems that I'm working on is schema inference. So Asterisk Database (AstDB) no database and we, and since schema is not ready as, as it would be for a SQL database, my duty was to create a schema for no SQL database.

Similarly, the next project would be about how do we improve the group by sort by query in the data in the AstDB. So my work online mostly works in the database kernel and how to improve the algorithm within a database.

1. What tools you use for managing retrieving, cleaning and analysis of the data?

So as part of my RA I haven't worked with data cleaning or managing the data quality of quality of the data that is being passed through the database. But as part of my earlier research projects, I, I've used we have used natural language processing to clean up the data from Twitter database, to ensure that we read the tokens are most important words are used because for example, in even though a lot of verbs, the words are there in the tweet, it is only after cleaning it and getting quantifiable values out of it that can be used for us for processing. So similar. So as part of the language or tweet that has been passed, we use natural language processing to delete all the store unnecessary words out of it and to ensure the keywords are required. Based on this, you either develop a batch pipeline or a stream pipeline. Apache Spark, Apache Spark Streaming, Scala, PySpark and Apache Beam are some of the most common Data Pipelining SDKs available for development of the ETL pipelines. So, as part of my early project, that's what I've done.

There are various ways of cleaning. For example, a lot of null data and Excel sheets out there that has to be removed. These are some of the edge cases, but I haven't worked directly on it. But these are my experience from the projects that I worked on.

1. What are the most common Data Warehousing use cases you see?

So, from my previous experience in a fintech startup called metamorphosis The there, there used to be a use case of collecting past insurance cases and these sort of since these are structured data and cannot change as, as opposed to the no SQL databases.

Since, this is a structure, there is a huge collection of data that is being generated from each insurance company and that is being stored in the data warehouse using the existing tools that is existing out there, but I haven't worked that on.

1. At what level of normal form you normalize data in your company?

So, the previous places that I worked on stuck on to third normal form because going for 4 normal form is redundant and would add an overhead that would not be agile as part of the agile development. So, third normal form is what I commonly see, but I haven't worked directly on designing the normal forms of the database, but this is what I've seen from my experience.

1. What are some of the key challenges that you see come up during the datawarehouse development process?

So, if the data that is being gene passed on to the data warehouse isn't processed or the quality of the data isn't good. Any analysis that comes out of that would be incorrect or would give a false direction to the company that is using the data. So, there's a lot of examples of using old data or archive data pass to just pass on to the data Barrows, which may not be useful.

Then another key challenges that is coming up is maintaining your own data warehouse as opposed to using the cloud. While there are a lot of benefits to it, there are also privacy challenges that has to be confirmed. For example, in my previous place that I work for. even though the companies are from Japan, Singapore and Europe, the privacy rules for each of these companies are different and managing those would be one of t I think.

So, one of the biggest challenges that comes up when you're building up data Barrows to ensure that the data is the privacy of the data is managed carefully and insight that we take from it. Also, it can be will be hindrance or a benefit based on the policies that have been followed from each of the countries. Another one is ensuring the raid or the replication of data to ensure that they are not lost during these processes.

As opposed to a large scale like Amazon who, who have their own skill set up there where throwing away or bring up a new, their hardest storage solution might not be as easier for a smaller scale data warehouse that is being generated in house. But these are some of the high overview of challenges that I faced but having haven't worked with this, This is the insight that I could provide with for this problem.

1. Have you ever used surrogate keys?

They, they are, they are primary keys that are uniquely identified, the record and usually their ID’s. But yeah, so that's how we use it. We haven't gone much more detail into that.

1. What challenge do you face while adding new data and how do you manage a growing database?

So, this is something that I have directly worked in the current research. The reason we did create a schema is to help database administrators to manage the data that is flowing into because no SQL databases can scale horizontally. And because of that, there's a chance that it could go out of control if the number of data sources are not managed or not all, not all API S that are being passed into a database is strictly cast with certain languages of the Java, but it could be loose as in javascript. So, this sort of horizontal scaling that does occur to give an insight is what something that the schema influence would do. And the so one, so as the number of in iot applications that the number of data points that could be generated from a sensor could vary from sensor to sensor. This sort of growing database has to be managed even though it's no SQL allows agile development.

It is also important to manage this sort of data that is in the database to put in perspective even though hard storage solutions are cheap. Twitter recently has deleted all of our prior to 2008 or 2000, 2018 or 2016 all data prior to that because storage solutions do have a cost to the company in operation cost.

And this is one of the growing challenges even though it's cheap there. It does incur a cost of storing all this data and especially as it grows. it's very important to manage it for the database administrator or the developer or the stakeholders to manage it properly.

1. What challenges you face when you have to merge 2 or more big database to an existing one?

I haven't faced any of such a situation. So I wouldn't be in the right place to give an answer for that.

1. Any recommendation or advise for someone who want to work in the data space?

So, one of the key interesting fields that I come across is the Amazon red ship architecture. And what's interesting is compared to the earlier read architectures. There is machine learning being used to identify when a storage is going to fulfil. So, using machine learning, they pre-emptively right, bring switch on a new storage solution to ensure that the data is not lost.

So, there's machine learning not being added to the solution right now. So that's an exciting space. Then the storage layer in the database cloud solution that is out there such as Snowflake data breaks. The storage layer is getting smarter and smarter. There is a layer of cashing that the data, the storage layer itself knows what to provide for the user.

But at the same time on the other side of the query, the horizontal scaling of the query, query, query modules of these databases are also getting smarter and smarter and have yet not been abstracted. So that is also an interesting space then the scaling of data. So that is one in terms of architecture that is one space that is really exciting when it comes to research.

Other is how do we handle large amount of data as we know that every year the number of devices and the data that is being generated is exponentially growing. And at the same time, the number of insights that we get from all these data is a very small percentage. So how do we manage unprocessed or unstructured data?

For example, how do we get insights from video and audio data sources? Because that is what's being produced a lot via TikTok or reels or whatever. How do we produce insights out of it? So, these are some of the areas that I find interesting. I'm sure there is a lot more in this field.

**Question 5 b:**

[**https://drive.google.com/drive/folders/1rgJ\_IKvcIGbeOpVhIcNwx7bxEwoYt\_4l?usp=sharing**](https://drive.google.com/drive/folders/1rgJ_IKvcIGbeOpVhIcNwx7bxEwoYt_4l?usp=sharing)