

<b>Course Outline</b>	<b>2</b>
Lectures	2
Syllabus	2
Textbooks	2
Evaluation	2
TAs	2
<b>Assignment 1 [Due on 14/11/2022]</b>	<b>3</b>
Instructions	3
Exercises	3
Answers to Google Classroom Posts	5

# Course Outline

## Lectures

Timetable: F slot in Auditorium

## Syllabus

[CSE-IITH UG Core Courses](#)

1. Database Introduction
2. Relational Model
3. Basic and Intermediate SQL
4. Formal-Relational Query Languages
5. Database Design using ER Model

## Textbooks

Course Textbook

[Database Systems Concepts, A. Silberschatz, H. Korth and S. Sudarshan, McGraw Hill, 7<sup>th</sup> Edition](#)

Others

[Database Management Systems, R. Ramakrishnan and J. Gehrke, 3<sup>rd</sup> Edition](#)

[Fundamentals of Database Systems, R. Elmasri and S. B. Navathe, Addison Wesley, 7<sup>th</sup> Edition](#)

[Database Systems: The Complete Book, H. Garcia, J. Ullman, J. Widom, 3<sup>rd</sup> Edition](#)

## Evaluation

Homework/Assignments: 50%

Exam/Quizzes: 50%

## TAs

Suryamukhi K <[cs17m19p100001@iith.ac.in](mailto:cs17m19p100001@iith.ac.in)>

## Assignment 1 [Due on 14/11/2022]

### Instructions

1. This assignment must be done individually.
2. Download and install SQLite from [here](#). You can see [here](#) for how to set up SQLite in Windows. Similarly you can do for other OS.
3. If we find cases of copying, then to all those who are involved we will be given either FR grade or 0 marks for the assignment. Please don't share your code or report with anyone.

### Exercises

In this assignment, you will design a relational database for storing information about the [National Basketball Association](#). You will be using a small subset of the data called [NBA dataset](#) in the form of multiple CSV files. These CSV files contain information about basketball players, teams, coaches, etc. You can download them from [here](#). Refer [here](#) for more detailed description of different attributes.

**[00 points] Exercise 1:** Create a database called nbadb.

**[10 points] Exercise 2:** The provided dataset has 8 CSV files. In each CSV file, the first row contains a list of attributes and the remaining rows contain the attribute values. You should create a relational table for each of the CSV files. For example, the file teams.txt contains the attributes: team, location, name, leag. You should create a table with a schema: team (team, location, name, leag), where the attributes (team, leag) together are the primary key.

You should specify all the necessary integrity constraints, such as NOT NULL, CHECK, UNIQUE, PRIMARY KEY, FOREIGN KEY, etc. Save all your DDL statements in a file **nba\_create\_table.sql**.

Note: While creating the schema, use the same attribute names as given in the CSV files. You can execute all the statements in a .sql file by using the command: `.read nba_create_table.sql`

#### Example:

```
CREATE TABLE teams (  
    team CHAR(3) NOT NULL,  
    location VARCHAR(20) NOT NULL,  
    name VARCHAR(20),  
    leag CHAR(1) NOT NULL,  
    PRIMARY KEY (team, leag));
```

**[05 points] Exercise 3:** Insert all the information given in the CSV files to the corresponding tables. Save the insert statements in a file named **nba\_load\_data.sql**. (Note: You need to skip the first line which has attribute names)

**[60 points] Exercise 4:** Write SQL queries for the following. Save all the queries in a single file named **nba\_query.sql**.

1. **[10 points]** Use table players to find the count of the number of players in each position.
2. **[10 points]** Find the top-5 most productive years, which is determined based on the total number of games played (gp) by all the players, including both regular seasons and playoffs, for each year. Solve ties by preferring chronologically older years, and print only the years.
3. **[10 points]** In the table player\_regular\_season\_career, add a new column eff (efficiency rating), which is defined as follows:
$$eff = (pts + reb + ast + stl + blk - ((fga - fgm) + (fta - ftm) + turnover))$$
Among the players who have played more than 500 games, find the top-10 most efficient players.
4. **[15 points]** Find the number of players who have played more regular season games in the year 1990 than regular season games in any other year in their career.
5. **[15 points]** Use table player\_regular\_season\_career to find the all-time best players. Use the two attributes gp (games played) and eff (efficiency rating) to compare players. For two players p1 and p2, we define that p1 dominates p2 if and only if p1 has a higher gp and eff value than p2.

Find a set of players (ilkid, firstname, lastname, gp, eff) P, so that each player in P is not dominated by any other player in the table player\_regular\_season\_career. Return the output in ascending order of ilkid.

**[5 points] Exercise 5:** Create a file called **nba\_delete\_all.sql** to drop all the tables, views, or other things you created in nbadb.

To restore to nbadb database content, you can use the following sequence of commands:

```
.read nba_delete_all.sql  
.read nba_create_table.sql  
.read nba_load_data.sql
```

What to submit?

1. nba\_create\_table.sql
2. nba\_load\_data.sql
3. nba\_query.sql

4. nba\_delete\_all.sql
5. Results.pdf: For the SQL queries in Exercise 4, you should submit the results in the file Results.pdf.
6. You should submit all your files as one zip file named as rollnumber.zip (e.g. cs21btech11001.zip) in Google classroom.

## Answers to Google Classroom Posts

1. Data cleaning
  - a. If the foreign key constraint is not being satisfied just due to some additional extra spaces added at the end of the primary key, then you need to trim the additional extra spaces using the TRIM function. First of all insert all the data using the .import function. Then update the values in the primary key attribute by trimming the extra spaces.
  - b. Make all the lowercase entries to uppercase entries. For e.g. smithch03 -> SMITHCH03
2. Constraints
  - a. Do not do any manual renaming of entries. Do not allow insert of rows that do not satisfy the foreign key constraint. This will result in less number of rows in some of your tables compared to the given csv files.
3. Query
  - a. In the file player\_regular\_season\_career.csv, some players have 2 entries. One for leag=N and another one for leag=A. You need to add both these rows for such players while computing efficiency rating, dominated players, etc.