**Interview Questions and Answers**

1) How would you quantify the influence of a Twitter user?

Ans first we can see what is the type of author. It can be  journalist, professional, authority, activist, investor, company or celebrity. Then other features of an account like number of the followers, number of likes and dislikes number of retweets of each tweets can be took into the consideration. If the data is not labelled on the basis of this features clustering can be done to group more influencing people.

2) Given data on congressional bills and which congressio- nal representatives co-sponsored the bills, how would you determine which other representatives are most similar to yours in voting behavior? How would you evaluate who is the most liberal? Most republican? Most bipartisan?                                                                                                               Ans : by seeing the proposed points of the congressional bills and the past history of congressional and whoever is more focusing on your needs can be choosed as my most simlar voting behaviour person.

3) Data Analysis       How would you come up with an algorithm to detect pla- giarism in online content?

Ans Most websites that detect plagerism simply rely on search engines such as Google, Yahoo and Bing to do the heavy lifting for them. The Google Advanced Search can be used to match an exact word or phrase, or this can simply be done by wrapping your search query in quotes.

Usually Plagerism detectors break down the entered document into phrases and then use search engines to find webpages that have matching phrases. if too many matching phrases, especially from the same source then your docuent is most likely plagerised. The phrases are usually the sentences, or chunks of sentences from the document .

4) Data Analysis       You have data on all purchases of customers at a grocery store. Describe to me how you would program an algo- rithm that would cluster the customers into groups. How would you determine the appropriate number of clusters to include?

Ans : The segmentation can be based on any type depends on the requirement one motive can be grouping the premium customers who give the most profit, in this case the total purchase (per month) of a customer can be taken into the consideration. Based on this feature customers can be segmented into premium or normal customer. The number of clusters totally depends on the problem domain knowledge, ex- if this problem want to focus only on premium customers then K should 2 one for premium customers another for non-premium customers. If one does not have domain knowledge then either through visualization the number of clusters can be decide or there are other methods too to find most optimal number of clusters like elbow method by making a graph between SSE and k we can decide the number of clusters.

5) Let’s say you’re building the recommended music engine at Spotify to recommend people music based on past lis- tening history. How would you approach this problem?                                                                                                                 Ans: most of the recommendation system rely on collaborative filtering. The idea of [collaborative filtering](http://en.wikipedia.org/wiki/Collaborative_filtering) is to **determine the users’ preferences from historical usage data**. For example, if two users listen to largely the same set of songs, their tastes are probably similar. Conversely, if two songs are listened to by the same group of users, they probably sound similar. This kind of information can be exploited to make recommendations. But they are content –agnostic, the same type of model can be used to recommend books, movies or music, for example.

6 ) Given that you have wifi data in your office, how would you determine which rooms and areas are underutilized and overutilized?

Ans:

1. Collect data: of wifi-accessed computers / laptops / cell phones, number in different time of a work day in each room
2. Calculate mean of number of accessed devices in each room, mean of online time in each room in a week or month
3. Generally the higher mean of number and time, the room is more utilized. We can roughly define underutilized as room with mean less than average, and vice versa
4. We still need to compare with the usage for each room, meeting room, office room or café room, compare the mean within rooms for the similar usage.

7) How could you use GPS data from a car to determine the quality of a driver?

Ans: To collect the data the vehicles can be equipped with an array of sensor, communication, and GPS technology to gather information on vehicle position, speed, position, yaw rate, heading, and time. . First, a simple table of GPS points—including such data items as latitude, longitude, speed, time of day, and heading—must be processed and translated into a trip-log format that describes travel behaviour in terms of a related set of origins, destinations, trips, routes   . Second, in order to construct relationships between travel behaviour and the surrounding built environment, the trip-log items must be associated with spatial features in GIS layers.  With the geographical bodies we can know that in which kind of areas what is the speed of the car ,it may be industrial area, or a highway or a residential area. According to some surveys on highways the old drivers drive fast, this kind of survey information can be used and compared with the GPS data of any driver and by making visualizations it can be analyse.

8) Given accelerometer, altitude, and fuel usage data from a car, how would you determine the optimum acceleration pattern to drive over hills?

Ans: Given the historical data of    accelerometer, altitude, and fuel usage data from a car, and acceleration pattern of a car we can train a neural network where the x= (accelerometer, altitude, fuel usage   ) and y = acceleration pattern, Now given x can be train against y , and y^ can be predicted.

9) Given position data of NBA players in a season’s games, how would you evaluate a basketball player’s defensive ability?                                                                                                                    Ans: given the statistics of player and the matches, number of the matches played by player, number of the wins with player, number of the wins without player, number of the goals of the player in the season , ratio of gaols and number of matches played by the player, positions played by player, each match’s position wise records can be collected. Then if the defensive ability is measured between 1 to 5 then based on the data the defensive ability can be classified between 1 to 5.

10)         Communication   How would you explain to a group of senior executives why data is important?

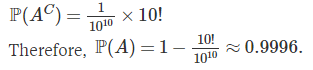
11)         Communication   How could you help the generate public understanding towards the importance of using data to generate in- sights?

12)         Communication   How would you convince a government agency to release their data in a publicly accessible API?

13)        Communication   I’m a local business owner operating a small restaurant. Convince me to switch my advertising budget from print to internet.

14)           Probability            Your hash function assigns each object to a number be- tween 1:10, each with equal probability. With 10 objects, what is the probability of a hash collision? What is the expected number of hash collisions? What is the expected number of hashes that are unused. .                                                                                                                  Ans:

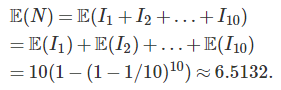
Let  A​ be the event of at least one hash collision. The probability of that event is equal to  1 − P().​ The complement is the event all hashes end up in different buckets. There are   ways of objects being hashed to numbers 1 through 10 and 10!​  ways the buckets could all be distinct. Assuming these are all equally likely, the probability is



Let H​ be the number of hash collisions. This is equal the number of *unused*buckets because for there to be a collision, then a bucket must be empty. Let N​ be the random variable that represents *used*buckets, so that H=10−N. The random variable N could be expressed as a sum of indicators. Denote by Ik​ the indicator that the integer 1≤k≤10 is *used*. Notice that for distinct indices, these are independent random variables. The expectation of I1​ is the probability that at least one of the objects get hashed to the bucket 1, call this event B1​ . This is equal to 1 minus the probability that none of the 10 objects get hashed to 1.



Now we use the linearity to get the expectation of *used*buckets



Therefore, the expected number of collisions is



15)          Statistical Inference            I have two different experiments that both change the sign-up button to my website. I want to test them at the same time. What kinds of things should I keep in mind?

16)          Statistical Inference            You are AirBnB and you want to test the hypothesis that a greater number of photographs increases the chances that a buyer selects the listing. How would you test this hypothesis?                                                                                                             Ans : For randomly selected listings with more than 1 pictures, we can hide 1 random picture for group A, and show all for group B. And can Compare the booking rate for the two groups. This can be way to test the hypothesis.

17)           Data Analysis       Given location data of golf balls in games, how would construct a model that can advise golfers where to aim?                                                                                                                                                                                                                                                        Ans : from the current location of the ball the ecludian distance can be calculated from each aim and the smallest distance can be choose to aim towords.

Previous 4 questions

**Q:Describe the difference between primary Key and foreign key in SQL DataBase?**

**Primary Key:**

A primary key is a field in a table which uniquely identifies each row/record in a database table. Primary keys must contain unique values. A primary key column cannot have NULL values.

A table can have only one primary key, which may consist of single or multiple fields. When multiple fields are used as a primary key, they are called a composite key.

**Foreign Key:**

A FOREIGN KEY is a key used to link two tables together.

A FOREIGN KEY is a field (or collection of fields) in one table that refers to the PRIMARY KEY in another table.

The table containing the foreign key is called the child table, and the table containing the candidate key is called the referenced or parent table.

| **BASIS FOR COMPARISON** | **PRIMARY KEY** | **FOREIGN KEY** |
| --- | --- | --- |
| Basic | Primary Key is a chosen candidate key that uniquely defines a tuple in a relation. | Foreign key in a table refers to the primary key of other table. |
| NULL | Primary key value can never be NULL. | Foreign key accepts NULL value. |
| Duplicate | No two tuples in a relation carry duplicate values for a primary key attribute. | Tuples can carry duplicate value for a foreign key attribute. |
| Temporary Table | Primary key constraint can be defined on the temporary tables. | Foreign Key constraint can not be defined on the temporary tables. |
| Insertion | We can insert a value to a primary key attribute, even if the referencing foreign key does not have that value in its column. | We can not insert a value to a foreign key, if that value is not present in the referenced primary key column. |
| Deletion | Before you delete a primary key value, make sure that value is not still present in the referencing foreign key column of referencing table. | You can delete a value from foreign key column without bothering, whether that value is present in referenced primary key column of referenced relation. |

**Code:-**

--Create Parent Table

CREATE TABLE Department

(

DeptID int PRIMARY KEY, --define primary key

Name varchar (50) NOT NULL,

Address varchar(100) NULL

)

GO

--Create Child Table

CREATE TABLE Employee

(

EmpID int PRIMARY KEY, --define primary key

Name varchar (50) NOT NULL,

Salary int NULL,

--define foreign key

DeptID int FOREIGN KEY REFERENCES Department(DeptID)

)

**Q: Given a COURSES table with columns course\_id and course\_name , a Faculty table with columns faculty\_id and faculty\_name,and a Course\_Faculty table with columns faculty\_id and course\_id ,how would you return a list of faculty who teach a course given the name of course?**

CREATE TABLE Courses (

course\_id int NOT NULL PRIMARY KEY,

course\_name varchar(30));

CREATE TABLE Course\_Faculty (

faculty\_id int NOT NULL PRIMARY KEY,

course\_id int references Courses(course\_id));

CREATE TABLE Faculty (

faculty\_id int references Course\_Faculty(faculty\_id),

faculty\_name varchar(30));

insert into Courses(course\_id,course\_name) values (340,'math');

insert into Courses(course\_id,course\_name) values (401,'bio');

insert into Courses(course\_id,course\_name) values (612,'physics');

insert into Course\_Faculty(course\_id,faculty\_id) values (340,190);

insert into Course\_Faculty(course\_id,faculty\_id) values (401,101);

insert into Course\_Faculty(course\_id,faculty\_id) values (612,200);

insert into Faculty(faculty\_name,faculty\_id) values ('Ram',190);

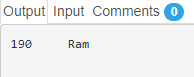
insert into Faculty(faculty\_name,faculty\_id) values ('Joker',101);

insert into Faculty(faculty\_name,faculty\_id) values ('Jack',612);

select faculty\_id,faculty\_name From Faculty

where faculty\_id= (Select faculty\_id from Course\_Faculty where course\_id =(select course\_id from Courses where course\_name="math"));

**OUTPUT:**

****

**Q:Write a query that returns the name of each department and a count of the number of employees in each :EMPLOYEES containing:Emp\_ID(Primary\_key) and Emp\_Name EMPLOYEES\_DEPT containing:Emp\_ID(Foreign key) and Dept\_ID(Foreign key) DEPT\_ID(Primary key) and Dept\_Name**

CREATE TABLE Employees (

emp\_id int NOT NULL PRIMARY KEY,

emp\_name varchar(30));

CREATE TABLE Depts (

dept\_id int NOT NULL PRIMARY KEY,

dept\_name varchar(30));

CREATE TABLE Employees\_dept (

emp\_id int references Employees(emp\_id),

dept\_id int references Depts(dept\_id));

insert into Employees(emp\_id,emp\_name) values (340,'jack');

insert into Employees(emp\_id,emp\_name) values (342,'joker');

insert into Employees(emp\_id,emp\_name) values (344,'jio');

insert into Depts(dept\_id,dept\_name) values (45,'Data science');

insert into Depts(dept\_id,dept\_name) values (55,'AI');

insert into Depts(dept\_id,dept\_name) values (65,'Deep Learning');

insert into Employees\_dept(dept\_id,emp\_id) values (45,340);

insert into Employees\_dept(dept\_id,emp\_id) values (65,342);

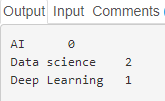
insert into Employees\_dept(dept\_id,emp\_id) values (45,344);

select Depts.dept\_name,count(emp\_id) From Depts

left join Employees\_dept ON Depts.dept\_id=Employees\_dept.dept\_id

group by Depts.dept\_name;

**OUTPUT:**



**Q: Given a IMPRESSION table with ad\_id,cliack(an indicator that the ad was clicked),and date,write a SQL query that will tell me the click-through-rate of each ad by month**

CREATE TABLE Impression

(

ad\_id INT ,

click bit,

dates date

);

insert into Impression(ad\_id,click,dates) values (340,0,'2019-03-12');

insert into Impression(ad\_id,click,dates) values (341,1,'2019-03-15');

insert into Impression(ad\_id,click,dates) values (343,0,'2019-03-15');

insert into Impression(ad\_id,click,dates) values (341,1,'2019-03-03');

insert into Impression(ad\_id,click,dates) values (343,1,'2019-03-1');

insert into Impression(ad\_id,click,dates) values (340,0,'2019-03-1');

insert into Impression(ad\_id,click,dates) values (340,1,'2019-03-05');

select ad\_id, average(click) from (select count(click) as click from Imprression group by ad\_id, month(dates)) group by ad\_id