**Normalization**

**Why do we need to do normalization?**  
  
To eliminate redundancy of data i.e. having same information stored at multiple places, which eventually be difficult to maintain and will also increase the size of our database.  
With normalization we will have tables with fewer columns which will make data retrieval and insert, update and delete operations more efficient.  
  
  
**What do we mean when we say a table is not in normalized form?**  
  
Let’s take an example to understand this,  
Say I want to create a database which stores my friends name and their top three favorite artists.  
This database would be quite a simple so initially I’ll be having only one table in it say friends table. Here FID is the primary key.

|  |  |  |
| --- | --- | --- |
| **FID** | **FNAME** | **FavoriteArtist** |
| 1 | Srihari | Akon, The Corrs, Robbie Williams. |
| 2 | Arvind | Enigma, Chicane, Shania Twain |

This table is not in normal form why?  
  
FavoriteArtist column is not atomic or doesn’t have scalar value i.e. it has having more that one value.  
Let’s modify this table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FID** | **FNAME** | **FavoriteArtist1** | **FavoriteArtist2** | **FavoriteArtist3** |
| 1 | Srihari | Akon. | The Corrs | Robbie Williams. |
| 2 | Arvind | Enigma | Chicane | Shania Twain |

This table is also not in normal form why?  
  
We have now changed our table and now each column has only one value!! (So what’s left?)  
Because here we are having multiple columns with same kind of value.  
  
I.e. **repeating group of data or repeating columns.**  
  
So what we need to do to make it normal or at least bring it in First Normal Form?

1. We’ll first break our single table into two.
2. Each table should have information about only one entity so it would be nice if we store our friend’s information in one table and his favorite artists’ information in another

(For simplicity we are working with few columns but in real world scenario there could be column like friend’s phone no, email , address and favorites artists albums, awards received by them, country etc. So in that case having two different tables would make complete sense)

|  |  |
| --- | --- |
| **FID** | **FNAME** |
| 1 | Srihari |
| 2 | Arvind |

|  |  |
| --- | --- |
| **FID** | **Favorite Artist** |
| 1 | Akon. |
| 1 | The Corrs |
| 1 | Robbie Williams |
| 2 | Enigma |
| 2 | Chicane |
| 2 | Shania Twain |

FID foreign key in FavoriteArtist table which refers to FID in our Friends Table.  
  
Now we can say that our table is in first normal form.  
  
Remember For **First Normal Form**  
  
**1...Column values should be atomic, scalar or should be holding single value**  
**2...No repetition of information or values in multiple columns.**  
  
**3...So what does Second Normal Form means?**  
  
  
 **Second normal form** our database should already be in first normal form and every non-key column must depend on entire primary key.  
  
Here we can say that our Friend database was already in second normal form l.  
Why?  
  
Because we don’t have composite primary key in our friends and favorite artists table.  
  
Composite primary keys are- primary keys made up of more than one column. But there is no such thing in our database.  
But still let’s try to understand second normal form with another example  
This is our new table

|  |  |  |  |
| --- | --- | --- | --- |
| **Gadgets** | **Supplier** | **Cost** | **Supplier Address** |
| Headphone | Abaci | 123$ | New York |
| Mp3 Player | Sagas | 250$ | California |
| Headphone | Mayas | 100$ | London |

In about table **ITEM+SUPPLIER** together form a composite primary key.  
  
Let’s check for dependency  
  
If I know gadget can I know the cost?  
  
No same gadget is provided my different supplier at different rate.  
  
If I know supplier can I know about the cost?  
  
No because same supplier can provide me with different gadgets.  
  
If I know both gadget and supplier can I know cost?  
  
Yes than we can.  
  
**So cost is fully dependent (functionally dependent) on our composite primary key (Gadgets+Supplier)**  
  
Let’s start with another non-key column Supplier Address.  
  
If I know gadget will I come to know about supplier address?  
  
Obviously no.  
  
If I know who the supplier is can I have it address?  
  
Yes.  
  
**So here supplier is not completely dependent on (partial dependent) on our composite primary key (Gadgets+Supplier).**  
  
This table is surely not in Second Normal Form.  
  
So what do we need to do to bring it in second normal form?  
  
Here again we’ll break the table in two.

|  |  |  |
| --- | --- | --- |
| **Gadgets** | **Supplier** | **Cost** |
| Headphone | Abaci | 123$ |
| Mp3 Player | Sagas | 250$ |
| Headphone | Mayas | 100$ |

|  |  |
| --- | --- |
| **Supplier** | **Supplier Address** |
| Abaci | New York |
| Sagas | California |
| Mayas | London |

We now how to normalize till second normal form.  
  
But let’s take a break over here and learn some definitions and terms.  
  
**Composite Key**: -Composite key is a primary key composed of multiple columns.  
**Functional Dependency** – When value of one column is dependent on another column.  
  
So that if value of one column changes the value of other column changes as well.  
  
e.g. Supplier Address is functionally dependent on supplier name. If supplier’s name is changed in a record we need to change the supplier address as well.  
  
S.Supplier–àS.SupplierAddress  
  
“In our s table supplier address column is functionally dependent on the supplier column”  
  
**Partial Functional Dependency –** A non-key column is dependent on some, but not all the columns in a composite primary key.  
  
In our above example **Supplier Address** was partially dependent on our composite key columns (Gadgets+Supplier).  
  
**Transitive Dependency-** A **transitive dependency** is a type of functional dependency in which the value in a non-key column is determined by the value in another non-key column.  
  
With these definitions in mind let’s move to Third Normal Form.  
**For a table in third normal form**

* It should already be in Second Normal Form.
* There should be no transitive dependency, i.e. we shouldn’t have any non-key column depending on any other non-key column.

Again we need to make sure that the non-key columns depend upon the primary key and not on any other non-key column.

|  |  |  |  |
| --- | --- | --- | --- |
| **Album** | **Artist** | **No. of tracks** | **Country** |
| Come on over | Shania Twain | 11 | Canada |
| History | Michael Jackson | 15 | USA |
| Up | Shania Twain | 11 | Canada |
| MCMXC A.D. | Enigma | 8 | Spain |
| The cross of changes | Enigma | 10 | Spain |

Although the above table looks fine but still there is something in it because of which we will normalize it further.  
  
Album is the primary key of the above table.  
  
Artist and No. of tracks are functionally dependent on the Album(primary key).  
  
But can we say the same of Country as well?  
  
In the above table Country value is getting repeated because of artist.  
  
So in our above table Country column is depended on Artist column which is a non-key column.  
  
So we will move that information in another table and could save table from redundancy i.e. repeating values of Country column.

|  |  |  |
| --- | --- | --- |
| **Album** | **Artist** | **No. of tracks** |
| Come on over | Shania Twain | 11 |
| History | Michael Jackson | 15 |
| Up | Shania Twain | 11 |
| MCMXC A.D. | Enigma | 8 |
| The cross of changes | Enigma | 10 |

|  |  |
| --- | --- |
| **Artist** | **Country** |
| Shania Twain | Canada |
| Michael Jackson | USA |
| Enigma | Spain |

**Normally this is considered enough and we don’t really go on applying the other normal forms.**   
  
**Most of real-world application has databases which are in third normal forms.**