Database Design, Implementation, and Normalization

Database designers use entity relationship diagrams to model databases. During the implementation phase the entity relationship diagrams are converted to a set of relational tables. Since the models are an abstraction of the real world applying normalization to the database may be required. The intent of normalizing a database is to reduce data redundancy and improve data integrity. Typically, a database is implemented in the third normal form and then tuned for performance. The cost of applying the normalization process can be minimized if we observe some basic rules while creating the entity relationship models.

Many database designers directly create the relation tables and then use a tool to generate the models. The tools that generate the models from the tables usually use the UML notation. A problem with the UML notation is that it may not be as easily interpreted by the customer. The entity relationship diagrams are visually easier and faster to understand.

Modeling the database allows us to:

- Communicate the desired structure and behavior of the system.
- Visualize and control the system's architecture.
- Better understand the system we are building.
- Manage risk.

Modeling the database accomplishes the following:

- Visualize the system as it is or as we want it to be.
- Permit us to specify the structure or behavior of the system.
- Provide a template that guides us in constructing the system.
- Document the decisions we have made.

Managing risk and documenting the decisions we have made are critical components in the system design and the need for it should not be under-estimated.

The entity relationship diagram displayed in Figure #1 models the relationship between an employee and associated project assignments. A high level analysis of the model raises several concerns.

Concern 1.
The relation schema Employee shows employee names as a single attribute instead of multiple attributes such as Last Name, First name, and Middle Initial. Showing an entity relationship diagram at a higher level of abstraction is not an unusual practice but requires a stronger interpretation when converting the diagram to tables.

Concern 2.
The relation schema Address is a weak entity. A primary key needs to be developed when translating the diagram into the relation tables. Introducing a foreign key SS# and using the attribute Type could form a composite primary key.

Concern 3.
The relation schema Project includes the attribute ProjectMgr. This raises a cautionary flag since the relation schema Employee contains the attribute Title. These two attributes may potentially contain duplication of data.

Concern 4.

The attribute ProjectID may not be an adequate primary key, since multiple project managers may be assigned to the project over different periods of time.

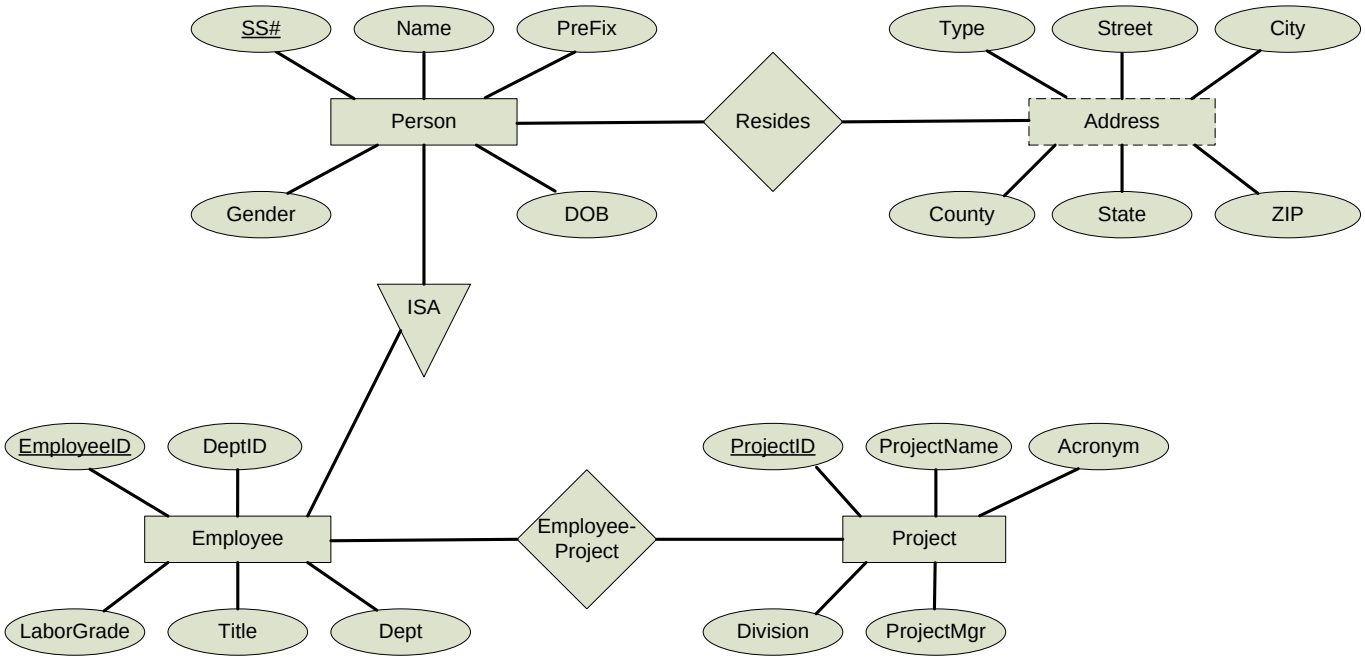

Figure #1.

A translation from the entity relationship diagram to database tables is shown below. Example data is added to the tables to show duplication of data, if any.

Person

SS#	Prefix	LastName	FirstName	MNI	DOB	Gender
220441111	Mr	Smith	Gregory	B	4/12/1985	M
330441111	Mrs	Jones	Elizabeth	S	7/21/1982	F
220441112	Mr	Carey	John	J	10/24/1982	M
330441112	Mrs	Blake	Nancy	null	11/3/1982	F
220441113	Mr	Hamilton	Fred	A	10/25/1960	M
220441114	Mr	Harrold	Thomas	B	11/15/1952	M
220441115	Mr	Messer	Bernt	K	1/1/1952	M
220441116	Mr	Goul	Josua	T	9/3/1956	M
220441117	Dr	Oligschlager	Albert	null	5/12/1963	M

Address

SS#	Type	Street	City	County	State	ZIP
220441111	Resident	12 Palm Street	Orlando	Orange	FL	32801
220441111	Shipping	PO Box 1200	Orlando	Orange	FL	32800
330441111	Resident	100 Oak Ave	Oviedo	Seminole	FL	32765
220441112	Resident	20 Palmetto Ave	Titusville	Brevard	FL	32780
330441112	Resident	71 Hibiscus Ave	Titusville	Brevard	FL	32780
220441113	Resident	120 Oak Ave	Oviedo	Seminole	FL	32766
220441114	Resident	19 SunFlower Street	Ovideo	Seminole	FL	32766
220441115	Resident	14 Rose Ave	Orlando	Orange	FL	32801
220441116	Resident	90 Hibiscus Ave	Titusville	Brevard	FL	32780
220441117	Resident	1024 SunFlower Street	Oviedo	Seminole	FL	32765

Employee

EmployeeID	SS#	LaborGrade	Title	DeptID	Dept
2501	220441111	L1	Junior Electrical Engineer	300	Software Development
2007	330441111	L3	Senior Test Engineer	250	Test Engineering
2005	220441112	L3	Senior Electrical Enginner	200	Electrical Engineering
2200	330441112	L2	Test Engineer	250	Test Engineering
2001	220441113	L4	Staff Software Engineer	300	Software Development
1020	220441114	L5	Sr Staff Accounting	100	Finance
1200	220441115	L5	Project Manager	150	Program Management
1001	220441116	L6	Project Manager	150	Program Management
1000	220441117	L6	Project Manager	150	Program Management

Project

EmployeeID	ProjectID	ProjectName	Acronym	Division	ProjectMgr
1200	1	Mission Moon	MM	Orlando	Harrold
1001	1	Mission Moon	MM	Orlando	Messer
1200	2	Deep Space Communications	DSC	Orlando	Goul
1000	2	Deep Space Communications	DSC	Orlando	Oligschlager
1001	3	Launch Control	LCDSC	Orlando	Goul

In this example the relationship Employee_Project was not translated into a separate table, instead table Project implements the relationship between Employee and Project by using a composite primary key. As we will realize soon this design implementation is not optimal. Following the database normalization process will help us understand the problems associated with the above implementation.

First Normal Form (1NF)

A relation schema is considered in First Normal Form when all attributes are atomic.
Each attribute must be indivisible and single valued.

An attribute may not be multi-valued or be comprised of lists and arrays.

Consider the relation schema Person.
The schema is comprised of seven attributes and SS# is the primary key.
All the attributes meet the requirements of 1NF, therefore relation schema Person is in 1NF.

Person

SS#	Prefix	LastName	FirstName	MNI	DOB	Gender

Consider the relation schema Address.
The schema is comprised of seven attributes and SS# and Type form the composite primary key.
The schema may be in 1NF if contents of the attribute Street is never decomposed into the street number and street name. Some database applications require that the street name and street numbers are stored separately. In this design the individual components of attribute Street can be combined and the relation schema Address is considered in 1NF.

Address

SS#	Type	Street	City	County	State	ZIP

Consider the relation schema Employee.
The schema is comprised of six attributes and EmployeeID is the primary key.
All the attributes meet the requirements of 1NF, therefore relation schema Employee is in 1NF.

Employee

EmployeeID	SS#	LaborGrade	Title	DeptID	Dept

Consider the relation schema Project.
The schema is comprised of six attributes. EmployeeID and ProjectID form the composite primary key. ProjectName, Acronym, Division, and ProjectMgr are all single valued attributes and meet the requirements for 1NF.

Project

EmployeeID	ProjectID	ProjectName	Acronym	Division	ProjectMgr

The functional dependencies for the database design are listed below.

FD1: SS# → Prefix, LastName, FirstName, MNI, DOB, Gender	(Person)
FD2: SS#, Type → Street, City, County, State, ZIP	(Address)
FD3: EmployeeID, → SS#, LaborGrade, Title, DeptID, Dept	(Employee)
FD4: EmployeeID, ProjectID → ProjectName, Acronym, Division, ProjectMgr	(Project)

Second Normal Form (2NF)

A relation schema is considered in Second Normal Form if it is in 1NF and if all non-key attributes are dependent on all attributes forming the primary key.

Since a partial dependency occurs when a non-key attribute is dependent on only a part of the composite key, the definition of 2NF is sometimes phrased as, "A table is in 2NF if it is in 1NF and if it has no partial dependencies."

Analyzing the functional dependencies shows that FD1 and FD3 have primary keys that are comprised of a single attribute. Therefore the relation schemas Person and Employee are in 2NF. All associated attributes are functionally dependent upon this single valued primary key.

Functional dependencies FD2 and FD4 have composite primary keys. The attributes associated with the key require evaluation to ensure that they depend on all of the components of the key.

Analyzing FD2 indicates that all attributes are dependent on the entire key, SS# and Type. A person can have multiple addresses, a primary residency, a work address, a mailing address, a work address, and a shipping address to name a few. The schema is in 2NF.

Analyzing the functional dependency FD4 shows that the relation schema Project is in 1NF but it is not in 2NF. The non-key attributes are not fully functionally dependent on the primary key. Attributes ProjectName, Acronym, and Division are functionally dependent on ProjectID, and ProjectMgr is functionally dependent on EmployeeID. FD4 requires decomposition into FD4 and FD5, resulting into the functional dependencies shown below.

FD1: SS# → Prefix, LastName, FirstName, MNI, DOB, Gender	(Person)
FD2: SS#, Type → Street, City, County, State, ZIP	(Address)
FD3: EmployeeID, → SS#, LaborGrade, Title, DeptID, Dept	(Employee)
FD4: EmployeeID → Title	(Employee-Project)
FD5: ProjectID → ProjectName, Acronym, Division	(Project)

We should now realize that FD4 describes the relationship Employee-Project in the Entity Relationship Diagram and the relationship between Employee and Project has cardinality of many-to-many. A many-to-many relationship cannot be simplified by omitting the relationship table.

Also, the relationship needs to show the duration of time when an employee was assigned to a particular project. It is possible for an employee to be assigned to a project at various times. The title of an employee may also change due to promotions or accepting different responsibilities working on different projects. Therefore the attribute Title in Employee and Employee-Project may not contain the same information. Employee-Project contains historical data. The functional dependencies and relation schemas are modified to reflect this change.

FD1: SS# → Prefix, LastName, FirstName, MNI, DOB, Gender	(Person)
FD2: SS#, Type → Street, City, County, State, ZIP	(Address)
FD3: EmployeeID, → SS#, LaborGrade, Title, DeptID, Dept	(Employee)
FD4: EmployeeID, ProjectID, From → Title, To	(Employee-Project)
FD5: ProjectID → ProjectName, Acronym, Division	(Project)

Third Normal Form (3NF)

A relation schema is considered in Third Normal Form (3NF) if it is in 2NF and the non-key attributes have no transitive dependencies.

Consider the functional dependency FD2.
Attributes City, County, and State have transitive dependencies on ZIP.
The Zip code uniquely identifies every region in the United States and the associated data redundancy can be reduced by eliminating the transitivity.

FD2: EmployeeID, Type → Street, ZIP → City, County, State

Consider the functional dependency FD3.
Attributes LaborGrade and Title have functional dependency on EmployeeID, but Dept has a transitive functional dependency on DeptID.

FD3: EmployeeID, → LaborGrade, Title, DeptID → Dept

The new set of functional dependencies is complying with 3NF. The entity relationship diagram is shown in Figure #2.

FD1: SS# → Prefix, LastName, FirstName, MNI, DOB, Gender	(Person)
FD2: SS#, Type → ZIP, Street	(Address)
FD3: ZIP → City, County, State	(Region)
FD4: EmployeeID, → SS#, LaborGrade, Title, DeptID	(Employee)
FD5: DeptID → Dept	(Department)
FD5: EmployeeID, ProjectID, From → Title, To	(Employee-Project)
FD6: ProjectID → ProjectName, Acronym, Division	(Project)

After transforming the relation schemas into 3NF an inspection is necessary to verify that all data (attributes) can still be obtained, usually by performing a natural join operation. The database designer must ensure that the functional decomposition is lossless and functional dependency preserving. Converting a database to 3NF should not present an issue; however if the database is converted to BCNF, functional dependency preservation is not guaranteed. Table 1 below shows that all data can be reached (obtained) given SS# in entity relation Person.

Person	**SS#**	Prefix	LastName	FirstName	MNI	DOB	Gender
Address	**SS#**	Type	Street	ZIP			
Region				**ZIP**	City	County	State
Employee	**SS#**	**EmployeeID**	LaborGrade	Title	DeptID		
Department					**DeptID**	Dept	
Employee-Project		**EmployeeID**	**ProjectID**	**From**	To	Title	
Project			**ProjectID**	ProjectName	Acronym	Division	

Table 1

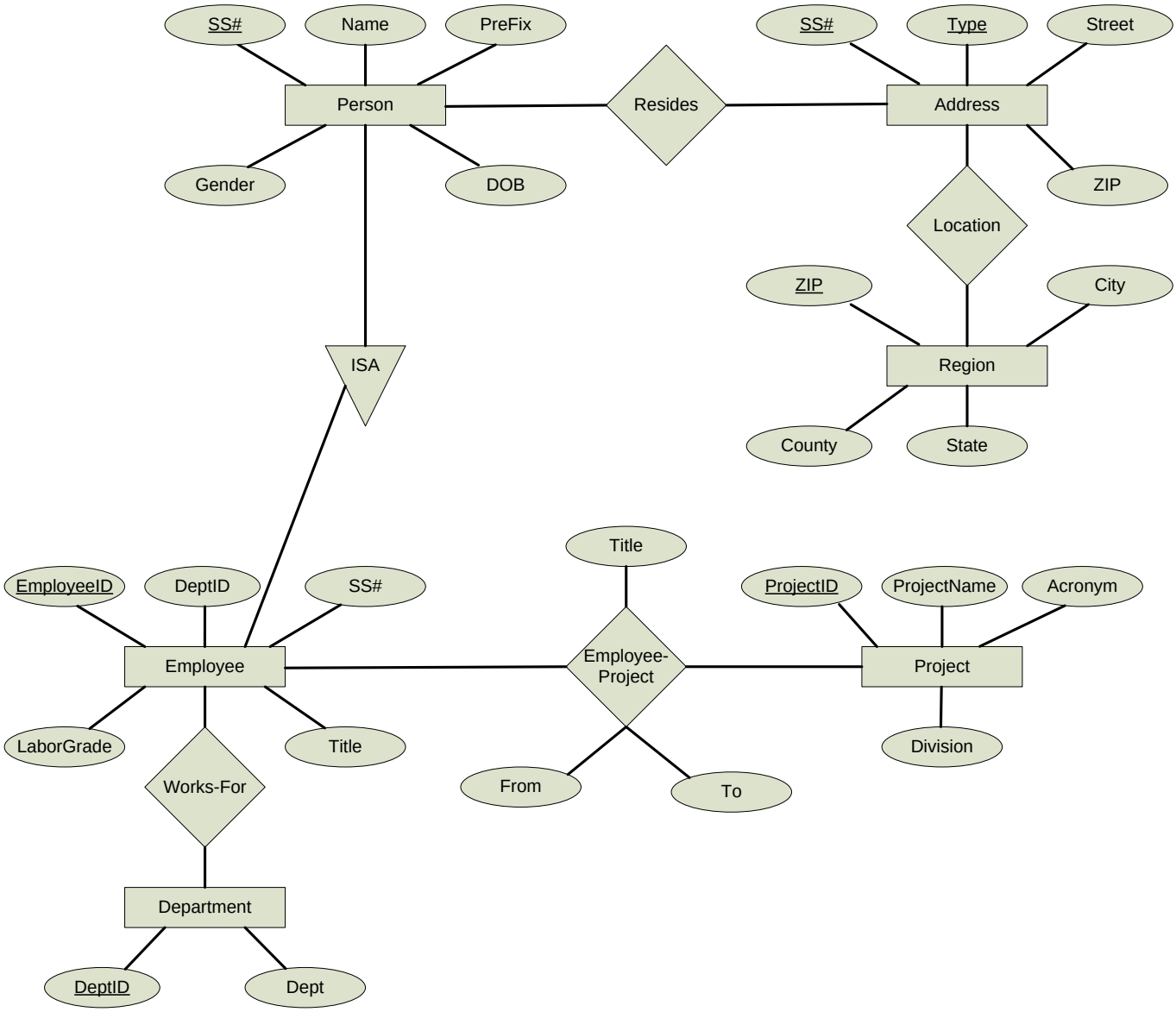
SS#
Name
PreFix
Person
Gender
DOB
Resides
SS#
Type
Street
Address
ZIP
Location
ZIP
City
Region
County
State
ISA
Title
EmployeeID
DeptID
SS#
Employee
LaborGrade
Works-For
Title
Department
DeptID
Dept
Employee-
Project
From
To
ProjectID
ProjectName
Acronym
Project
Division

Figure #2

Person

SS#	Prefix	LastName	FirstName	MNI	DOB	Gender
220441111	Mr	Smith	Gregory	B	4/12/1985	M
330441111	Mrs	Jones	Elizabeth	S	7/21/1982	F
220441112	Mr	Carey	John	J	10/24/1982	M
330441112	Mrs	Blake	Nancy	null	11/3/1982	F
220441113	Mr	Hamilton	Fred	A	10/25/1960	M
220441114	Mr	Harrold	Thomas	B	11/15/1952	M
220441115	Mr	Messer	Bernt	K	1/1/1952	M
220441116	Mr	Goul	Josua	T	9/3/1956	M
220441117	Dr	Oligschlager	Albert	null	5/12/1963	M

Address

SS#	Type	Street	ZIP
220441111	Resident	12 Palm Street	32801
220441111	Shipping	PO Box 1200	32800
330441111	Resident	100 Oak Ave	32765
220441112	Resident	20 Palmetto Ave	32780
330441112	Resident	71 Hibiscus Ave	32780
220441113	Resident	120 Oak Ave	32766
220441114	Resident	19 SunFlower Street	32766
220441115	Resident	14 Rose Ave	32801
220441116	Resident	90 Hibiscus Ave	32780
220441117	Resident	1024 SunFlower Street	32765

Region

ZIP	City	County	State
32801	Orlando	Orange	FL
32800	Orlando	Orange	FL
32765	Oviedo	Seminole	FL
32780	Titusville	Brevard	FL
32780	Titusville	Brevard	FL
32766	Oviedo	Seminole	FL
32766	Ovideo	Seminole	FL
32801	Orlando	Orange	FL
32780	Titusville	Brevard	FL
32765	Oviedo	Seminole	FL

Employee

EmployeeID	SS#	LaborGrade	Title	DeptID
2501	220441111	L1	Junior Electrical Engineer	300
2007	330441111	L3	Senior Test Engineer	250
2005	220441112	L3	Senior Electrical Enginner	200
2200	330441112	L2	Test Engineer	250
2001	220441113	L4	Staff Software Engineer	300
1020	220441114	L5	Sr Staff Accounting	100
1200	220441115	L5	Project Manager	150
1001	220441116	L6	Project Manager	150
1000	220441117	L6	Project Manager	100

Department

DeptID	Dept
100	Finance
150	Program Management
200	Electrical Engineering
250	Test Engineering
300	Software Development

Employee-Project

EmployeeID	ProjectID	From	To	Title
1200	1	12/1/1980	12/30/1995	Project Manager
1001	1	1/1/1990	2/12/1998	Project Manager
1200	2	1/2/1996	7/30/2010	Project Manager
1000	2	2/15/1998	7/30/2010	Project Manager
1001	3	2/13/1998	7/30/2010	Project Manager

Project

ProjectID	ProjectName	Acronym	Division
1	Mission Moon	MM	Orlando
1	Mission Moon	MM	Orlando
2	Deep Space Communications	DSC	Orlando
2	Deep Space Communications	DSC	Orlando
3	Launch Control	LCDSC	Orlando