Chapter 4
MORE NORMALIZATION

Introduction

In this fourth chapter, we will explore going beyond effective data modeling to include a new perspective on the normalization process. Normalization can be overwrought with rules and formulas; the goal in this chapter is to simplify this critical process and enable readers a better —and quicker—method for designing effective data structures that can scale without the burden of having to learn arcane rules.

During this simplification we will expound on the horizontal design approach introduced in the previous chapter and reinforce the need to recognize relationship patterns. Hopefully, each reader will gain insight and confidence to construct fully normalized database structures quickly by using the standard practice presented here. Normalization is just putting pieces of a puzzle together.

Structure

By the end of this chapter, readers will be able to conduct the following:

- Articulate two different approaches to normalization
- Understand the basics of what makes a relation and the violations that render any ERD into Unnormalized Form (UNF)
- Feel more comfortable in the process of normalization and be able to identify a design in a specific normal form from UNF to 4NF
- Recognize the patterns that most object relationships hold as well as how to further normalize a particular design
- Become a design leader capable of taking a collection of functional requirements and creating a professional data model in at least 3NF

A Closer Look at Normalization

An apt comparison for designing a data model is that it resembles making music; sure, there is the science of scales and key signatures behind every hit song but many people with zero musical training still recognize good music when they hear it without knowing exactly why. Most people can recognize dissonant notes because they are jarring to the ear. A common goal with the normalization process is that we trust our eyes when we see columns that are misplaced like certain combinations of notes that simply do not sound pleasant when played together.

The purpose of normalization is to construct a data repository that not only holds the data necessary for an organization to capture and analyze ("learn"), but also to do so at volume with guarantees that the data is pristine and trustworthy. Converting the knowledge we have obtained about a business into an effective data model requires more than passing awareness of the normalization process. It is imperative that all data modelers recognize not only poor database designs on-site but also become aware of common mistakes included in many flawed databases. If a database is not fully normalized, there are several complications that arise, which lead to a combination of poor performance as well as inconsistent (a fancy term also known as 'bad') or missing data.

Two Methods of Conducting Normalization Process

Pick up a random book about database normalization and it might feel like a 9th grade calculus class staring at equations with Greek symbols. That might be the science behind a 'well designed' database, but that is not how the best designers think. The best database designers advocate for a more 'musical' approach that hopefully resonates with most readers when compared to the traditional 'mathy' method.

Quickly, the two methods of normalization referred to in this textbook are as follows:

- Relational calculus ('mathy')
- Ironing ('musical')

Ironing is a process that resembles "ironing out the wrinkles" of a design where columns that have made it through brainstorming happen to be temporarily located in the wrong entity. This ironing process is the preferred method for many data science students when creating original structures for class projects. This method will be emphasized shortly in a step-by-step process of designing the METRO_TRANSIT database.

Let's get into the specific definitions of the different normal forms all the way from unnormalized to fourth normal form (4NF) with some examples. First, however, we must start with defining a *relation*. In relational theory, a relation is defined as a collection of values that have the same characteristics. We have several names for these, including relation, entity, and finally a table. Please see the table below for when we change titles:

Phase	**Name for the data Collection**	**Name for an element of a Collection**	**Name for a single instance of values**
Conceptual	Relation	Field	Record
Logical	Entity	Attribute	Tuple
Physical	Table	Column	Row

Table 4.1 Labels used for the names of objects based on design phase

While many professionals will confuse the names of objects in the list above, it is important for all technicians to be aware of the formal names for each object and when to reference their use. Most developers will refer to the objects listed above by the name associated with them in the physical design phase.

Now, before we become more familiar with the various normal forms, we must address the list of violations that will prevent the establishment of a relation (also known as a 'table').

Violation	**Description**
Non-unique relation name	The name of the relation must be unique within the database
Non-unique attribute name	Within a relation, each attribute must have unique name
Non-unique row	Within a relation, each row must be unique (we solve this with primary key)
Multi-valued attribute	Each cell of an attribute can contain only a single value
Inconsistent domain value	Each attribute must have a single data type and a range of values
Row-dependency	The order of rows shall not matter (unlike when people make lists)

Table 4.2: Violations that will render a design to be labeled Unnormalized Form (UNF)

Unnormalized Form (UNF)

Unnormalized form is the state of an entire relational database if any violation of the definition of a relation is present. Better said is if any object in the database that will hopefully become a table has any violations from Table 4.2 above, the entire database is said to be in Unnormalized Form or simply UNF.

Let's see an example of UNF with sample data placed in a spreadsheet/table from METRO_TRANSIT; this example has many issues that

DateTime	Passenger	RouteNumber/Name	Type	Driver	StopName	Type	Destination	Type	Fare	Neighborhood
2/13/2025 7:36:00 AM (delayed)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix, 206555-7661	Broadway Ave Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
2/13/2025 9:36	Darcel Eustache	78 Fremont-Waterfront-Downtown	Special	Meryl Streep	Hwy 99-N 36th	Covered	Fremont, Space Needle, P70	Extra-long	$ 2.75	Fremont
2/14/2025 6:36		78 Fremont-Waterfront-Downtown	XP	Bruce Lee	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
2/16/2025 6:32	Kenyetta Terron	42 Sodo-Downtown Express	Express	Meryl Streep	First Avenue and Terry Street	Covered	Fremont, Space Needle, P70	Doub	$ 2.75	SODO
2/19/2025 15:39	Kenny Terron	42 - E Sodo-Downtown Express	Espress	Bruce Lee	First Avenue and Terry Street	Uncovered	SoDo, DT	Doubled	$ 4.50	South Downtown
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 7:36	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
2/22/2025 8:32	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jim Morrison	Broadway Ave and Cherry St	Covered	CCntr,	Articulated	$ 4.50	Capital Hill
3/1/2025 6:33			Regular	Bruce Lee, 425 6109225	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/1/2025 6:36:00 AM (Blocked)			Regular	Mary Streep (mstre@mtransit.org)	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/3/2025 6:42	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	Jim Morrison	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/5/2025 6:32			XP	Meryl Streep	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/9/2025 7:52	Kenyetta Terron	42 Sodo-Downtown Express	Express	Jimi Hendrix	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
3/10/2025 6:33	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	jim	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/11/2025 6:32	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee; blee@metrotran.org	First Avenue and Terry Street	Uncovered	Downtown	Doubled	$ 4.50	South Downtown
3/13/2025 8:32:00 AM (Late)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Bruce Lee	Broadway Ave and Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
3/14/2025 6:32			Express	Bruce Lee	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	C Hill
3/15/2025 1:36	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Jimmy Hendricks	Broadway Ave and Cherry St	Covered	Convention Center	2 Decks	$ 2.75	Capitol Hill
3/16/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Jim Morrison	Sixth Ave and Battery Street	Regular	Fremont, Space Needle	Doubled	$ 4.50	Downtown
3/17/2025 11:33	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix	Fourth Avenue and Seneca	Covered	Capitol Hill	Doubled	$ 2.75	Downtown
3/18/2025 16:13	Jane Lundgran	32 Capitol Hill-Downtown	Reg	Meryl Streep	Fourth Avenue and Seneca St	Covered	CH	2 Decks	$ 2.75	DT
3/19/2025 6:32	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Meryl Streep	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	Capitol Hill
3/21/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Meryl Streep	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/26/2025 9:04	Janie Lundgren	32 Capitol Hill-Downtown	Regular	Jimmy Hendricks, 206 5557661	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	CHill
4/18/2025 5:23	Darcie Eustache	78 Fremont-Waterfront-Downtown	Commute	Jimi Hendrix	Fourth Avenue and Seneca	Regular	Fremont, Pier 70, Needle	Doubled	$ 4.50	Downtown

Figure 4.1: Example set of data on METRO_TRANSIT tracking BOARDINGS in UNF

The data in Figure 4.1 shows how a person might track METRO_TRANSIT in a spreadsheet; this is considered an unnormalized form or UNF for reasons to be shown in the next several sections. Please accept that while this is a very common method of organizing data, it will be proven to be flawed!

Many people will try to capture data in the shape of how they will be reading it (as if it is a report). This is a result of centuries of practical human organization or categorization of objects. The design and structure of storing data (on disk or in the cloud) is going to be different from how it will be read by people.

Violation of a Relation #1: Non-unique relation/entity name

These examples of each violation from Table 4.2 are included so we can see each violation up close and in slow motion. For the first violation, having two relations ('tables') with the same name, is the same problem as duplicate file names in the same folder on a computer. There is no 'screen shot' to show as an image here, just trust that each entity must be uniquely named to be coded into a relational database management system.

Violation of a Relation #2: Non-unique attribute name

DateTime	Passenger	RouteNumber/Name	Type	Driver	StopName	Type	Destination	Type	Fare	Neighborhood
2/13/2025 7:36:00 AM (delayed)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix, 206555-7661	Broadway Ave Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
2/13/2025 9:36	Darcel Eustache	78 Fremont-Waterfront-Downtown	Special	Meryl Streep	Hwy 99-N 36th	Covered	Fremont, Space Needle, P70	Extra-long	$ 2.75	Fremont
2/14/2025 6:36		78 Fremont-Waterfront-Downtown	XP	Bruce Lee	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
2/16/2025 6:32	Kenyetta Terron	42 Sodo-Downtown Express	Express	Meryl Streep	First Avenue and Terry Street	Covered	Fremont, Space Needle, P70	Doub	$ 2.75	SODO
2/19/2025 15:39	Kenny Terron	42 - E Sodo-Downtown Express	Espress	Bruce Lee	First Avenue and Terry Street	Uncovered	SoDo, DT	Doubled	$ 4.50	South Downtown
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 7:36	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
2/22/2025 8:32	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jim Morrison	Broadway Ave and Cherry St	Covered	CCntr,	Articulated	$ 4.50	Capital Hill
3/1/2025 6:33			Regular	Bruce Lee, 425 6109225	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/1/2025 6:36:00 AM (Blocked)			Regular	Mary Streep (mstre@mtransit.org)	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/3/2025 6:42	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	Jim Morrison	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/5/2025 6:32			XP	Meryl Streep	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/9/2025 7:52	Kenyetta Terron	42 Sodo-Downtown Express	Express	Jimi Hendrix	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
3/10/2025 6:33	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	jim	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/11/2025 6:32	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee; blee@metrotran.org	First Avenue and Terry Street	Uncovered	Downtown	Doubled	$ 4.50	South Downtown
3/13/2025 8:32:00 AM (Late)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Bruce Lee	Broadway Ave and Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
3/14/2025 6:32			Express	Bruce Lee	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	C Hill
3/15/2025 1:36	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Jimmy Hendricks	Broadway Ave and Cherry St	Covered	Convention Center	2 Decks	$ 2.75	Capitol Hill
3/16/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Jim Morrison	Sixth Ave and Battery Street	Regular	Fremont, Space Needle	Doubled	$ 4.50	Downtown
3/17/2025 11:33	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix	Fourth Avenue and Seneca	Covered	Capitol Hill	Doubled	$ 2.75	Downtown
3/18/2025 16:13	Jane Lundgran	32 Capitol Hill-Downtown	Reg	Meryl Streep	Fourth Avenue and Seneca St	Covered	CH	2 Decks	$ 2.75	DT
3/19/2025 6:32	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Meryl Streep	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	Capitol Hill
3/21/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Meryl Streep	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/26/2025 9:04	Janie Lundgren	32 Capitol Hill-Downtown	Regular	Jimmy Hendricks, 206 5557661	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	CHill
4/18/2025 5:23	Darcie Eustache	78 Fremont-Waterfront-Downtown	Commute	Jimi Hendrix	Fourth Avenue and Seneca	Regular	Fremont, Pier 70, Needle	Doubled	$ 4.50	Downtown

Figure 4.2: Example violation of a relation with non-unique attributes in a spreadsheet

In the set of data found in Figure 4.2, please observe there are three different columns with an identical name of TYPE; it appears that these are intended to track types for ROUTE, STOP, and DESTINATION. A relational database management system will not allow these duplications as it is a violation of the definition of a relation.

Recognize that a human *might* be able to decipher this flaw in labeling; we often text each other in shorthand or abbreviations. A computer or database management system on the other hand will be interpreting labels in a literal fashion and will not be able to discern the subtle difference displayed here in Figure 4.2.

Violation of a Relation #3: Non-unique row

DateTime	Passenger	RouteNumber/Name	Type	Driver	StopName	Type	Destination	Type	Fare	Neighborhood
2/13/2025 7:36:00 AM (delayed)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix, 206555-7661	Broadway Ave Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
2/13/2025 9:36	Darcel Eustache	78 Fremont-Waterfront-Downtown	Special	Meryl Streep	Hwy 99-N 36th	Covered	Fremont, Space Needle, P70	Extra-long	$ 2.75	Fremont
2/14/2025 6:36		78 Fremont-Waterfront-Downtown	XP	Bruce Lee	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
2/16/2025 6:32	Kenyetta Terron	42 Sodo-Downtown Express	Express	Meryl Streep	First Avenue and Terry Street	Covered	Fremont, Space Needle, P70	Doub	$ 2.75	SODO
2/19/2025 15:39	Kenny Terron	42 - E Sodo-Downtown Express	Espress	Bruce Lee	First Avenue and Terry Street	Uncovered	SoDo, DT	Doubled	$ 4.50	South Downtown
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 7:36	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
2/22/2025 8:32	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jim Morrison	Broadway Ave and Cherry St	Covered	CCntr,	Articulated	$ 4.50	Capital Hill
3/1/2025 6:33			Regular	Bruce Lee, 425 6109225	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/1/2025 6:36:00 AM (Blocked)			Regular	Mary Streep (mstre@mtransit.org)	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/3/2025 6:42	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	Jim Morrison	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/5/2025 6:32			XP	Meryl Streep	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/9/2025 7:52	Kenyetta Terron	42 Sodo-Downtown Express	Express	Jimi Hendrix	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
3/10/2025 6:33	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	jim	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/11/2025 6:32	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee; blee@metrotran.org	First Avenue and Terry Street	Uncovered	Downtown	Doubled	$ 4.50	South Downtown
3/13/2025 8:32:00 AM (Late)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Bruce Lee	Broadway Ave and Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
3/14/2025 6:32			Express	Bruce Lee	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	C Hill
3/15/2025 1:36	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Jimmy Hendricks	Broadway Ave and Cherry St	Covered	Convention Center	2 Decks	$ 2.75	Capitol Hill
3/16/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Jim Morrison	Sixth Ave and Battery Street	Regular	Fremont, Space Needle	Doubled	$ 4.50	Downtown
3/17/2025 11:33	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix	Fourth Avenue and Seneca	Covered	Capitol Hill	Doubled	$ 2.75	Downtown
3/18/2025 16:13	Jane Lundgran	32 Capitol Hill-Downtown	Reg	Meryl Streep	Fourth Avenue and Seneca St	Covered	CH	2 Decks	$ 2.75	DT
3/19/2025 6:32	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Meryl Streep	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	Capitol Hill
3/21/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Meryl Streep	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/26/2025 9:04	Janie Lundgren	32 Capitol Hill-Downtown	Regular	Jimmy Hendricks, 206 5557661	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	CHill
4/18/2025 5:23	Darcie Eustache	78 Fremont-Waterfront-Downtown	Commute	Jimi Hendrix	Fourth Avenue and Seneca	Regular	Fremont, Pier 70, Needle	Doubled	$ 4.50	Downtown

Figure 4.3: Example violation of a relation with non-unique rows

The second example of a violation that creates a database relation to be labeled as UNF is when there are duplicate rows in a single relation. In the image from Figure 4.3, there are two identical rows highlighted in yellow. While these can be assumed to be erroneous, they are able to be recorded in a spreadsheet because a spreadsheet does not reject duplicate rows by default. This is a violation as duplicate rows in a relational database would be impossible to determine which row is being queried or referenced.

A relational database management system prevents duplicates because at least one column (the primary key) is guaranteed to be unique. The data shown in Figure 4.3 is in an example of Unnormalized Form or UNF.

Violation of a Relation #4: Multi-valued attribute

Next, we will see an example that shows a common violation of a relation with multiple values in a single cell. Please note that this error resembles violation #5 *Inconsistent domain value range*. The differences will be explained shortly.

DateTime	Passenger	RouteNumber/Name	Type	Driver	StopName	Type	Destination	Type	Fare	Neighborhood
2/13/2025 7:36:00 AM (delayed)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix, 206555-7661	Broadway Ave Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
2/13/2025 9:36	Darcel Eustache	78 Fremont-Waterfront-Downtown	Special	Meryl Streep	Hwy 99-N 36th	Covered	Fremont, Space Needle, P70	Extra-long	$ 2.75	Fremont
2/14/2025 6:36		78 Fremont-Waterfront-Downtown	XP	Bruce Lee	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
2/16/2025 6:32	Kenyetta Terron	42 Sodo-Downtown Express	Express	Meryl Streep	First Avenue and Terry Street	Covered	Fremont, Space Needle, P70	Doub	$ 2.75	SODO
2/19/2025 15:39	Kenny Terron	42 - E Sodo-Downtown Express	Espress	Bruce Lee	First Avenue and Terry Street	Uncovered	SoDo, DT	Doubled	$ 4.50	South Downtown
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 7:36	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
2/22/2025 8:32	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jim Morrison	Broadway Ave and Cherry St	Covered	CCntr,	Articulated	$ 4.50	Capital Hill
3/1/2025 6:33			Regular	Bruce Lee, 425 6109225	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/1/2025 6:36:00 AM (Blocked)			Regular	Mary Streep (mstre@mtransit.org)	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/3/2025 6:42	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	Jim Morrison	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/5/2025 6:32			XP	Meryl Streep	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/9/2025 7:52	Kenyetta Terron	42 Sodo-Downtown Express	Express	Jimi Hendrix	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
3/10/2025 6:33	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	jim	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/11/2025 6:32	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee; blee@metrotran.org	First Avenue and Terry Street	Uncovered	Downtown	Doubled	$ 4.50	South Downtown
3/13/2025 8:32:00 AM (Late)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Bruce Lee	Broadway Ave and Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
3/14/2025 6:32			Express	Bruce Lee	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	C Hill
3/15/2025 1:36	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Jimmy Hendricks	Broadway Ave and Cherry St	Covered	Convention Center	2 Decks	$ 2.75	Capitol Hill
3/16/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Jim Morrison	Sixth Ave and Battery Street	Regular	Fremont, Space Needle	Doubled	$ 4.50	Downtown
3/17/2025 11:33	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix	Fourth Avenue and Seneca	Covered	Capitol Hill	Doubled	$ 2.75	Downtown
3/18/2025 16:13	Jane Lundgran	32 Capitol Hill-Downtown	Reg	Meryl Streep	Fourth Avenue and Seneca St	Covered	CH	2 Decks	$ 2.75	DT
3/19/2025 6:32	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Meryl Streep	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	Capitol Hill
3/21/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Meryl Streep	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/26/2025 9:04	Janie Lundgren	32 Capitol Hill-Downtown	Regular	Jimmy Hendricks, 206 5557661	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	CHill
4/18/2025 5:23	Darcie Eustache	78 Fremont-Waterfront-Downtown	Commute	Jimi Hendrix	Fourth Avenue and Seneca	Regular	Fremont, Pier 70, Needle	Doubled	$ 4.50	Downtown

Figure 4.4: Example violation of a relation with multi-valued attributes

The fourth violation of the rules for avoiding UNF is known as a multi-valued attribute. In the image for Figure 4.4, the highlighted data represents multiple destinations listed in a single cell. For instance, in the yellow highlighted cells, there are multiple values in each cell, with DateTime (column 1) having descriptions of status as well as date and time data. The column of RouteNumber/Name (column 3) not only has multiple values in a single cell, it also has multiple bits of different data (RouteNumber like '32' but also the Name of the Route, 'Capital Hill-Downtown'. Driver (column 5) has multiple values that are also of different ranges, with names as well as contact information (phone or email).

Please note these examples are of two different violations of the definition of a relation!

- multiple values in any cell is a violation
- having different types/ranges of data in the same column is a second violation

Additional violations of having multiple values in a single cell that are highlighted in yellow in Figure 4.4 include DESTINATION.

While people can read these values and deduce that there are two or three distinct values in a single cell, a database management system is going to read this as a single value that has just a bunch more characters in its name (as well as commas!). Please also notice that

there are multiple occurrences of typos and misspellings; creating additional entities to manage these values will go a long way towards preventing alternate spellings of all values whether intentional or accidental.

Objects that have multiple values per relationship are very common in life, and we need to be able to accommodate them in our database designs. These are almost always a 'many-to-many' relationship and the best solution is creating two new entities, one for the column that has multiple values and a second for the associative entity that will be between these two entities. More on this solution is coming up!

Violation of a Relation #5: Inconsistent domain value range

DateTime	Passenger	RouteNumber/Name	Type	Driver	StopName	Type	Destination	Type	Fare	Neighborhood
2/13/2025 7:36:00 AM (delayed)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix, 206555-7661	Broadway Ave Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
2/13/2025 9:36	Darcel Eustache	78 Fremont-Waterfront-Downtown	Special	Meryl Streep	Hwy 99-N 36th	Covered	Fremont, Space Needle, P70	Extra-long	$ 2.75	Fremont
2/14/2025 6:36		78 Fremont-Waterfront-Downtown	XP	Bruce Lee	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
2/16/2025 6:32	Kenyetta Terron	42 Sodo-Downtown Express	Express	Meryl Streep	First Avenue and Terry Street	Covered	Fremont, Space Needle, P70	Doub	$ 2.75	SODO
2/19/2025 15:39	Kenny Terron	42 - E Sodo-Downtown Express	Espress	Bruce Lee	First Avenue and Terry Street	Uncovered	SoDo, DT	Doubled	$ 4.50	South Downtown
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 7:36	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
2/22/2025 8:32	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jim Morrison	Broadway Ave and Cherry St	Covered	CCntr,	Articulated	$ 4.50	Capital Hill
3/1/2025 6:33			Regular	Bruce Lee, 425 6109225	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/1/2025 6:36:00 AM (Blocked)			Regular	Mary Streep (mstre@mtransit.org)	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/3/2025 6:42	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	Jim Morrison	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/5/2025 6:32			XP	Meryl Streep	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/9/2025 7:52	Kenyetta Terron	42 Sodo-Downtown Express	Express	Jimi Hendrix	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
3/10/2025 6:33	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	jim	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/11/2025 6:32	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee; blee@metrotran.org	First Avenue and Terry Street	Uncovered	Downtown	Doubled	$ 4.50	South Downtown
3/13/2025 8:32:00 AM (Late)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Bruce Lee	Broadway Ave and Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
3/14/2025 6:32			Express	Bruce Lee	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	C Hill
3/15/2025 1:36	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Jimmy Hendricks	Broadway Ave and Cherry St	Covered	Convention Center	2 Decks	$ 2.75	Capitol Hill
3/16/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Jim Morrison	Sixth Ave and Battery Street	Regular	Fremont, Space Needle	Doubled	$ 4.50	Downtown
3/17/2025 11:33	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix	Fourth Avenue and Seneca	Covered	Capitol Hill	Doubled	$ 2.75	Downtown
3/18/2025 16:13	Jane Lundgran	32 Capitol Hill-Downtown	Reg	Meryl Streep	Fourth Avenue and Seneca St	Covered	CH	2 Decks	$ 2.75	DT
3/19/2025 6:32	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Meryl Streep	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	Capitol Hill
3/21/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Meryl Streep	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/26/2025 9:04	Janie Lundgren	32 Capitol Hill-Downtown	Regular	Jimmy Hendricks, 206 5557661	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	CHill
4/18/2025 5:23	Darcie Eustache	78 Fremont-Waterfront-Downtown	Commute	Jimi Hendrix	Fourth Avenue and Seneca	Regular	Fremont, Pier 70, Needle	Doubled	$ 4.50	Downtown

Figure 4.5: Example violation of a relation with domain violations in two columns

In Figure 4.5, the yellow-highlighted cells represent three columns that each have different kinds of data; this is problematic for database systems to try and decipher. Under column 'DateTime' there are at least two data ranges, the first is in fact datetime values, however the second range of values appears to be related to a status of some sort. 'Delayed', 'Blocked', and 'Late' are values that are distinct from datetime and while this might be able to be understood by a human, it is not able to be included in a relational database in this shape. Data stored in this manner will not be able to be stored in a database as there are different data types and objects being referenced.

As mentioned previously, in the column titled ‘Driver’, there are also several distinct ranges of values that are not a name. In this column, there appears to be phone numbers, and email data in addition to combined first and last names. Again, while people are used to interpreting multiple ranges in columns like these, database management systems will not.

We will see solutions to this flaw and how to improve the database design to be accommodated in a relational system.

Violation of a Relation #6: Row-dependency

DateTime	Passenger	RouteNumber/Name	Type	Driver	StopName	Type	Destination	Type	Fare	Neighborhood
2/13/2025 7:36:00 AM (delayed)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix, 206555-7661	Broadway Ave Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
2/13/2025 9:36	Darcel Eustache	78 Fremont-Waterfront-Downtown	Special	Meryl Streep	Hwy 99-N 36th	Covered	Fremont, Space Needle, P70	Extra-long	$ 2.75	Fremont
2/14/2025 6:36		78 Fremont-Waterfront-Downtown	XP	Bruce Lee	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
2/16/2025 6:32	Kenyetta Terron	42 Sodo-Downtown Express	Express	Meryl Streep	First Avenue and Terry Street	Covered	Fremont, Space Needle, P70	Doub	$ 2.75	SODO
2/19/2025 15:39	Kenny Terron	42 - E Sodo-Downtown Express	Espress	Bruce Lee	First Avenue and Terry Street	Uncovered	SoDo, DT	Doubled	$ 4.50	South Downtown
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 7:36	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
2/22/2025 8:32	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jim Morrison	Broadway Ave and Cherry St	Covered	CCntr,	Articulated	$ 4.50	Capital Hill
3/1/2025 6:33			Regular	Bruce Lee, 425 6109225	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/1/2025 6:36:00 AM (Blocked)			Regular	Mary Streep (mstre@mtransit.org)	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/3/2025 6:42	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	Jim Morrison	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/5/2025 6:32			XP	Meryl Streep	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/9/2025 7:52	Kenyetta Terron	42 Sodo-Downtown Express	Express	Jimi Hendrix	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
3/10/2025 6:33	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	jim	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/11/2025 6:32	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee; blee@metrotran.org	First Avenue and Terry Street	Uncovered	Downtown	Doubled	$ 4.50	South Downtown
3/13/2025 8:32:00 AM (Late)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Bruce Lee	Broadway Ave and Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
3/14/2025 6:32			Express	Bruce Lee	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	C Hill
3/15/2025 1:36	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Jimmy Hendricks	Broadway Ave and Cherry St	Covered	Convention Center	2 Decks	$ 2.75	Capitol Hill
3/16/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Jim Morrison	Sixth Ave and Battery Street	Regular	Fremont, Space Needle	Doubled	$ 4.50	Downtown
3/17/2025 11:33	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix	Fourth Avenue and Seneca	Covered	Capitol Hill	Doubled	$ 2.75	Downtown
3/18/2025 16:13	Jane Lundgran	32 Capitol Hill-Downtown	Reg	Meryl Streep	Fourth Avenue and Seneca St	Covered	CH	2 Decks	$ 2.75	DT
3/19/2025 6:32	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Meryl Streep	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	Capitol Hill
3/21/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Meryl Streep	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/26/2025 9:04	Janie Lundgren	32 Capitol Hill-Downtown	Regular	Jimmy Hendricks, 206 5557661	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	CHill
4/18/2025 5:23	Darcie Eustache	78 Fremont-Waterfront-Downtown	Commute	Jimi Hendrix	Fourth Avenue and Seneca	Regular	Fremont, Pier 70, Needle	Doubled	$ 4.50	Downtown

Figure 4.6: Example violation of a relation with row dependencies

The 6th example of a violation of the rules for constructing a relation is what is called a ‘row dependency’ and is highlighted in Figure 4.6 above. A row dependency is where a row in a table requires an adjacent row (usually the row above) to provide context or values to be fully understood. For example, in Figure 4.6, we see several examples of the passenger and RouteName columns being empty with no values. When looking at this spreadsheet, a person may deduce that the blank cells are relying on the rows immediately preceding the empty cells for context. Looking at the second and third rows, we read passenger ‘Darcel Eustache’ quite likely is the intended value for the third row, which is blank.

While the third-row value for 'Passenger' may have been left blank on purpose, a computer system is not going to know which other surrounding rows or values are intended to be used for describing them. This may work when writing notes on the front of the fridge for your roommates, but it fails when trying to construct a relational data model and renders the entire database to be UNF if it is present.

Basic Requirements of Relation: Going from UNF to 1NF

Now that we have a more solid high-level understanding of what makes a relation, let us take a systematic approach of converting the non-normalized design of METRO_TRANSIT to the design of a relational database in first normal form or 1NF.

As mentioned earlier, the basic requirements for a relation are as follows:

- Each entity within a database must have a unique name.
- Each attribute (column) within an entity must also have a unique name.
- No row may depend on any other row for its existence ("row dependency").
- Each cell must contain at most one value (known as a "multivalued attribute").
- Each cell in a column must adhere to a single range of values (known as a "domain").

We are assuming that each element of data present as column headers is important and considered critical to the organization; we will work hard to find a more appropriate location for the values as opposed to just dropping it from the database.

First Normal Form (1NF)

A database design is automatically in First Normal Form (1NF) if it meets all the conditions for being a relation. This should be a beginning goal and not an end goal; some of the worst database designs I have witnessed in my life were at least 1NF and completely ineffective. We will get an opportunity to see METRO_TRANSIT progress from UNF to 3NF after getting an overview of all normal forms.

Second Normal Form (2NF)

A database design is automatically in Second Normal Form (2NF) if it meets all the conditions for being in 1NF and does not have what are called “partial dependencies”. This is where attributes in an entity describe only part of the primary key. Partial dependencies occur when designers try to combine several attributes to establish a “natural” primary key without resorting to a surrogate key (this is another argument for assigning surrogate keys.

Here is a very quick example of a partial dependency and a design in 2NF using the entity EMPLOYEE from an early design of METRO_TRANSIT with several other attributes added:

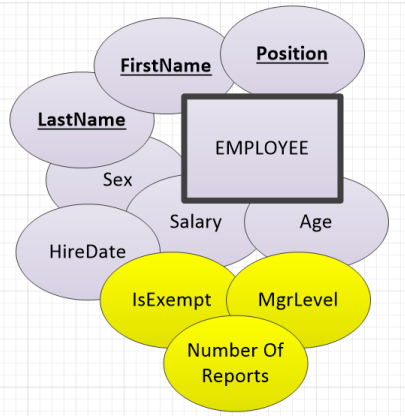


Figure 4.7: Example of a relation with partial dependencies

The entity EMPLOYEE in Figure 4.7 has three attributes underlined with bold type to indicate the composite primary key. The attributes IsExempt, MgrLevel, and NumberOfReports have been added to help with this explanation. The new attributes IsExempt and MgrLevel both describe the attribute Position, which is only part of the primary key. The attribute NumberOfReports describes only the combined attributes of FirstName and LastName, which is the other part of the primary key. Two different partial dependencies in one example!

This design, which includes a partial dependency, renders the entire database to First Normal Form or 1NF. The following image in Figure 4.8 is the design fix that places the same entity in 2NF as it no longer has any partial dependencies:

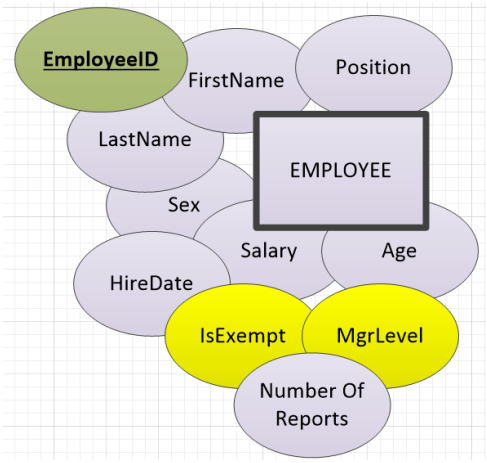


Figure 4.8: Example of a relation without a partial dependency

In Figure 4.8, there has been an addition of a surrogate primary key as opposed to having three attributes as a composite primary key. With this slight change, the attributes IsExempt, MgrLevel, and NumberOfReports no longer describe part of the primary key. While IsExempt and MgrLevel still both describe the attribute Position, since there is a surrogate key, this is called a transitive dependency as opposed to partial dependency.

Since NumberOfReports describes EmployeeID (the primary key), it is no longer problematic or cause of concern. Briefly, a transitive dependency is when an attribute is describing a non-key attribute. Breaking this down to a math formula of "A > B > C".

- Let 'A' represent the primary key
- Let 'B' represent a non-key attribute
- Let 'C' represent another attribute that is dependent on the value of the non-key

In Figure 4.8, 'A' is EmployeeID, 'B' is Position, and 'C' is both IsExempt by itself, and MgrLevel by itself (two instances of a transitive dependency). We often describe a transitive dependency as "when A defines B, and B defines C, then A transitively defines C". This is commonly written as "A > B > C". This table in Figure 4.8 is in second normal form or 2NF because it has a transitive dependency (technically two of them).

Third Normal Form (3NF)

A design is defined as being in 3NF if it meets all the requirements for 2NF and does not have any transitive dependencies. The resolution of the example in Figure 4.9 requires several new entities, including POSITION, POSITION_TYPE, and JOB:

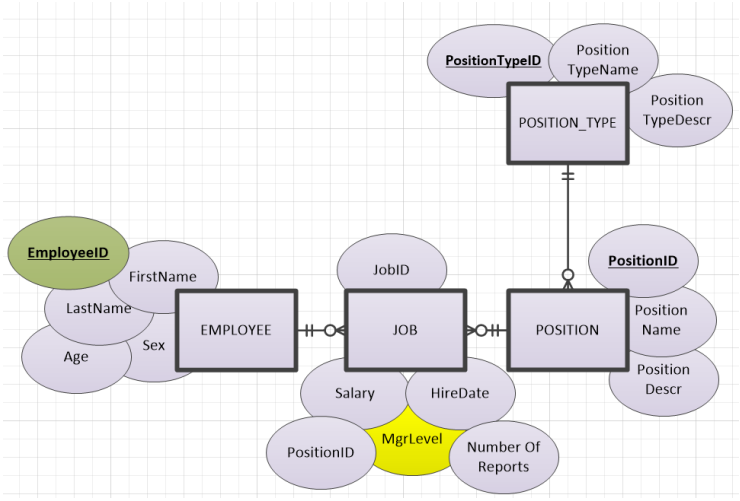


Figure 4.9: Example of a set of relations without a transitive dependency

In the example from Figure 4.9, there are no longer the transitive dependencies that were present in Figure 4.8. The entity of JOB is an associative entity resolving the M:M relationship between POSITION and EMPLOYEE as one employee over time can hold many positions within the same company or organization. Conversely, the same position (such as 'Shift Manager') can be held by many people over the lifetime of the database. Since an instance of JOB is one person and one position coming together, we can attach the attributes that were misplaced previously, as they are describing the position in the context of the person in the job. This provides flexibility in having a range of values for MgrLevel for example, where a new hire in the position of Shift Manager may have a Mgr Level of '1' while someone else with more experience will have a MgrLevel of '3', even though the PositionName is the same. As for the attribute 'IsExempt', it can be placed as a value under the new entity POSITION_TYPE and the attribute of PositionTypeName.

Fourth Normal Form (4NF)

A relational database design is defined as being in 4NF if it meets all the requirements for 3NF and does not have any multi-valued dependencies.

Multi-valued dependencies are found in the relationships that have NOT been fully articulated to be M:M even in rare cases. In Chapter 3 *Data Modeling and Normalization*, we saw several examples of a Consolidated Conceptual Diagram data model for METRO_TRANSIT. In that previous chapter, looking at Figure 3.23, there are only five associative entities defined. In Figure 3.24 (after a bit more analyses), there were an additional three defined for a total of eight. It can be argued that Figure 3.23 is in Third

Normal Form and Figure 3.24 is in Fourth Normal Form because all potential multi-valued dependencies were resolved. We will see more examples of the differences between 3NF and 4NF in the next section.

Hopefully, this brief foray into normalization has not made your interest in data science wane in any way. Again, one of the main objectives of this book is to enable every reader to be entirely independent of other data professionals; not relying on someone else to design a database for you because you have that skill already.

Normalization is a heavy lift to fully understand; very few people get to an expert level in this space for a reason. You have done a tremendous job getting through these weeds so far; pay attention to anything that is still somewhat confusing, and we can try and address your concerns with further practice or explanations. Now we move into the next section where we will go through the normalization process using a method that combines aspects of both the conceptual and logical design phases as we 'iron' the database into the desired shape of third or even fourth normal form.

Method 2: 'Ironing' a database from UNF to 4NF

The ironing method of building an Entity Relationship Diagram has a 'horizontal' nature as we emphasize a quick and shallow definition of entities while quickly defining relationships as we add more entities. This process might be more common if we have pre-existing systems, are in a hurry, or we are not as committed to proper documentation as in method 1. We should end up in the same spot as the first method eventually, so the preferred method will be whichever makes the most sense to you and your organization.

We have seen what the UNF version of METRO_TRANSIT looks like in Figure 4.1 (shown again for quick reference below):

DateTime	Passenger	RouteNumber/Name	Type	Driver	StopName	Type	Destination	Type	Fare	Neighborhood
2/13/2025 7:36:00 AM (delayed)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix, 206555-7661	Broadway Ave Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
2/13/2025 9:36	Darcel Eustache	78 Fremont-Waterfront-Downtown	Special	Meryl Streep	Hwy 99-N 36th	Covered	Fremont, Space Needle, P70	Extra-long	$ 2.75	Fremont
2/14/2025 6:36		78 Fremont-Waterfront-Downtown	XP	Bruce Lee	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
2/16/2025 6:32	Kenyetta Terron	42 Sodo-Downtown Express	Express	Meryl Streep	First Avenue and Terry Street	Covered	Fremont, Space Needle, P70	Doub	$ 2.75	SODO
2/19/2025 15:39	Kenny Terron	42 - E Sodo-Downtown Express	Espress	Bruce Lee	First Avenue and Terry Street	Uncovered	SoDo, DT	Doubled	$ 4.50	South Downtown
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 6:13	Darcel Eustache	78-S Fremont-Waterfront-Downtown	Special	Jimmy Hendricks, 206 5557661	Elliott Avenue and Mercer St	Cvd	Fremont, Space Needle, P70	Extra-long	$ 2.75	Interbay
2/21/2025 7:36	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
2/22/2025 8:32	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jim Morrison	Broadway Ave and Cherry St	Covered	CCntr,	Articulated	$ 4.50	Capital Hill
3/1/2025 6:33			Regular	Bruce Lee, 425 6109225	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/1/2025 6:36:00 AM (Blocked)			Regular	Mary Streep (mstre@mtransit.org)	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	Capitol Hill
3/3/2025 6:42	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	Jim Morrison	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/5/2025 6:32			XP	Meryl Streep	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/9/2025 7:52	Kenyetta Terron	42 Sodo-Downtown Express	Express	Jimi Hendrix	First Avenue and Terry Street	UC	Downtown	Doubled	$ 4.50	South Downtown
3/10/2025 6:33	Darcel Eustache	78 Fremont-Downtown Commuter	Commute	jim	Elliott Avenue and Mercer St	Covered	Downtown	Extra-long	$ 2.75	Interbay
3/11/2025 6:32	Kenyetta Terron	Sodo-Downtown Express	Express	Bruce Lee; blee@metrotran.org	First Avenue and Terry Street	Uncovered	Downtown	Doubled	$ 4.50	South Downtown
3/13/2025 8:32:00 AM (Late)	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Bruce Lee	Broadway Ave and Cherry St	Covered	Convention Center	Articulated	$ 4.50	Capitol Hill
3/14/2025 6:32			Express	Bruce Lee	Broadway Ave and Cherry St	Covered	Downtown	Articulated	$ 4.50	C Hill
3/15/2025 1:36	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Jimmy Hendricks	Broadway Ave and Cherry St	Covered	Convention Center	2 Decks	$ 2.75	Capitol Hill
3/16/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Jim Morrison	Sixth Ave and Battery Street	Regular	Fremont, Space Needle	Doubled	$ 4.50	Downtown
3/17/2025 11:33	Ivey Hazekamp	32 Capitol Hill-Downtown	Regular	Jimi Hendrix	Fourth Avenue and Seneca	Covered	Capitol Hill	Doubled	$ 2.75	Downtown
3/18/2025 16:13	Jane Lundgran	32 Capitol Hill-Downtown	Reg	Meryl Streep	Fourth Avenue and Seneca St	Covered	CH	2 Decks	$ 2.75	DT
3/19/2025 6:32	Janey Lundgren	32 Capitol Hill-Downtown	Reg	Meryl Streep	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	Capitol Hill
3/21/2025 6:32	Darcel Eustache	78 Fremont-Waterfront-Downtown	Commute	Meryl Streep	Sixth Avenue and Battery	Regular	Fremont, Space Needle, P70	Doubled	$ 4.50	Downtown
3/26/2025 9:04	Janie Lundgren	32 Capitol Hill-Downtown	Regular	Jimmy Hendricks, 206 5557661	Broadway and Cherry St	Covered	Downtown	2 Decks	$ 2.75	CHill
4/18/2025 5:23	Darcie Eustache	78 Fremont-Waterfront-Downtown	Commute	Jimi Hendrix	Fourth Avenue and Seneca	Regular	Fremont, Pier 70, Needle	Doubled	$ 4.50	Downtown

Reprint of Figure 4.1: Review of METRO_TRANSIT in UNF

To continue this exercise of taking a collection of data from UNF to perhaps 3NF involves using a data modeling tool. All the images reflecting a data modeling tool in this textbook are cut from Microsoft Visio 2010, which is no longer supported by Microsoft. This is a truly remarkable data modeling tool for roughly $20 USD if you can find a copy on the internet.

Let's take the spreadsheet data set from Figure 4.1 above and place it in a data modeling tool to begin our ironing exercise, it will look like the image in Figure 4.10 below:

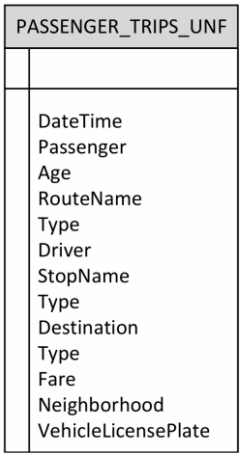


Figure 4.10: Example design of METRO_TRANSIT in UNF

Please note there are several reasons why this image is considered UNF, including all violations of what makes a 'relation'. Please refer to the images in Figures 4.2 through 4.9 earlier in this chapter for reminders on the following violations:

- Non-unique column names
- Row dependency
- Domain violation
- Multivalued attribute

Next, let's consider what the fixes to the violations will do to the design in Figure 4.1 (which again is in UNF). These fixes will need to address each of the violations specified in Table 4.2 to create a relation in 1NF. Please consider the design in Figure 4.11 which no longer has any violations outlined for a basic relation.

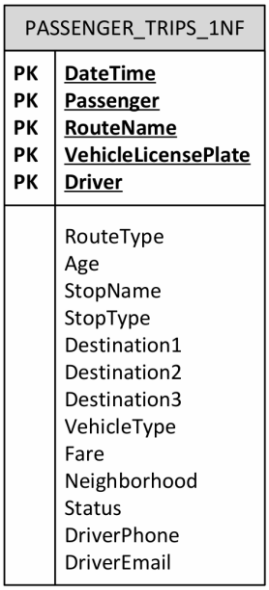


Figure 4.11 Example design of METRO_TRANSIT in 1NF

Remember, First Normal Form has met the requirements of a relation as none of the 6 violations are present, but there are still instances of other violations that keep the design in 1NF. These include what have been defined as 'partial dependencies'. In Figure 4.11 above, RouteType describes the attribute of RouteName, which obviously is only part of the three attributes that comprise the primary key. Likewise, all three Destination attributes also are describing only RouteName.

Other partial dependencies found in Figure 4.11 include the following:

- Age is describing Passenger
- VehicleTypeID describing VehicleLicensePlate
- DriverPhone and DriverEmail are both describing Driver

The presence of these attributes keeps the overall design in Figure 4.11 in 1NF. Next, let's address the partial dependency issue and see if there is an easy way to get the design into Second Normal Form (2NF). We can quickly substitute the composite (complicated?) primary key with a surrogate key that is intended to provide a unique value for each row.

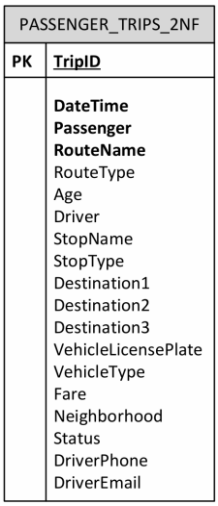


Figure 4.12: Example design of METRO_TRANSIT in 2NF

Observing Figure 4.12 above has a surrogate primary key that has replaced the previous natural composite primary key that had five attributes. Nothing else has changed in this design and immediately it is now 2NF. Please note this is still an abomination of a design, but for illustrative purposes the design meets the conditions of 2NF. Each of the previous partial dependencies that made the design 1NF are now considered 'transitive' dependencies. As we have read previously, the presence of these transitive dependencies prevents the design from being in Third Normal Form (3NF).

The next step is a bit tricky, as to take an overtly poor design (like the one in Figure 4.12) that is technically in 2NF into a design in 3NF, we must identify each attribute that can survive independently as an entity. This will take a few concerted steps outlined in the next few pages and includes a bit more brainstorming.

Please consider the diagram in Figure 4.13, which shows an extraction of rows from the single-table design in Figure 4.12 into a beginning collection of 13 entities.

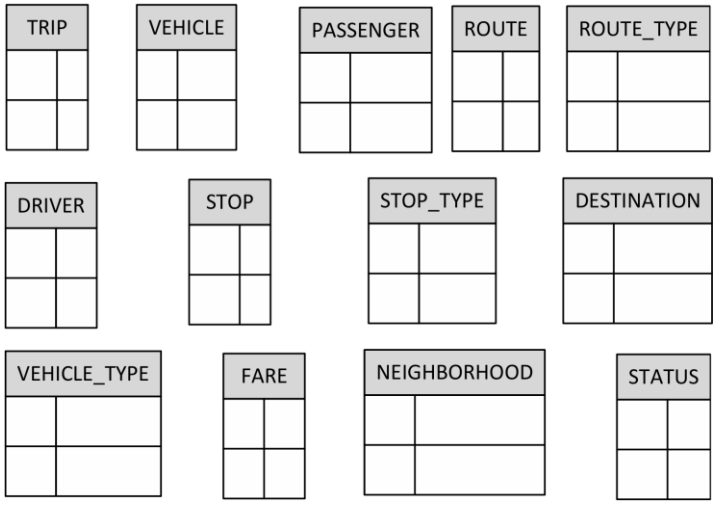


Figure 4.13: Identifying each potential new entity based on columns that are nouns

In the image of Figure 4.13, we can see the creation of 12 new entities based on the attributes specified in PASSENGER_TRIPS_2NF to make a total of 13. The next step in the ironing process includes adding surrogate primary keys to each entity.

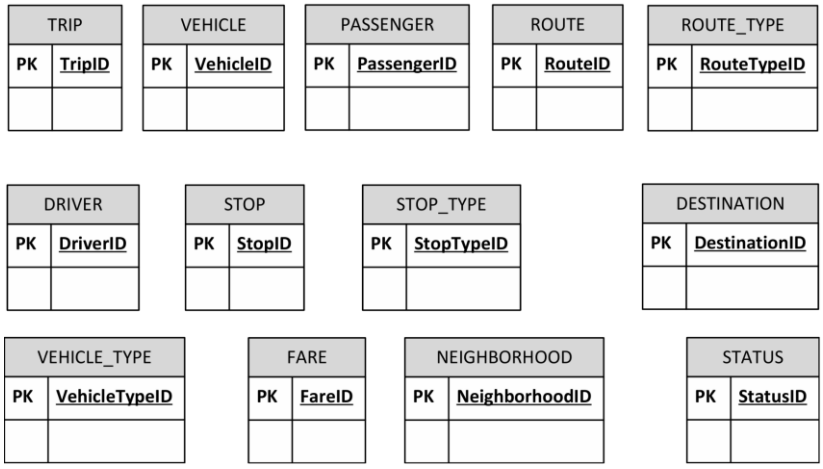


Figure 4.14: Adding surrogate PK values following tablename + ID

For now, a good choice is following 'tablename + ID' as the default name of the new primary key columns for each entity with a data type of INTEGER. This design provides flexibility for many different scenarios and allows best-practices such as auto-incrementing PK and physical indexes.

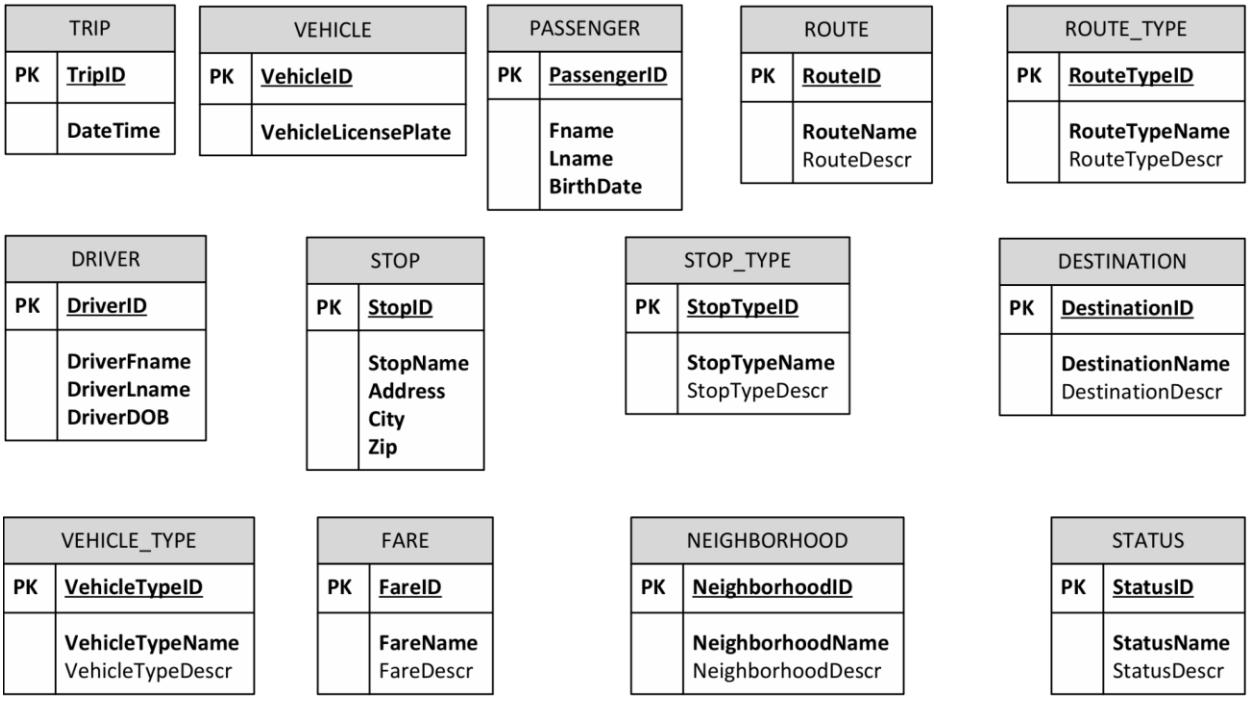


Figure 4.15: Adding Name, Description, and other obvious columns

Database Design Patterns

There are certain patterns that occur when following the 'ironing' method of database design. These include establishing surrogate primary keys as well as both a 'name' and 'description' attribute almost by default. There may be modification of this pattern on a case-by-case basis, but when we follow this template, it reduces the mistakes that are easily made when designers get in a hurry trying to over-burden an entity with attributes that belong somewhere else.

For now, trust that the process of ironing a database design has a track record of providing a straight-forward direction that leads to better-than-average outcomes with completed database designs that are fully normalized and scale well. A great achievement!

Next step in the design process is to take a high-level view of the entire diagram and begin to see which objects or entities are related to each other. Positioning related entities together at this stage will go a long way towards establishing a more easily read ERD and enable designers a more clear picture to consider as they continue with the normalization process by adding relationships.

.

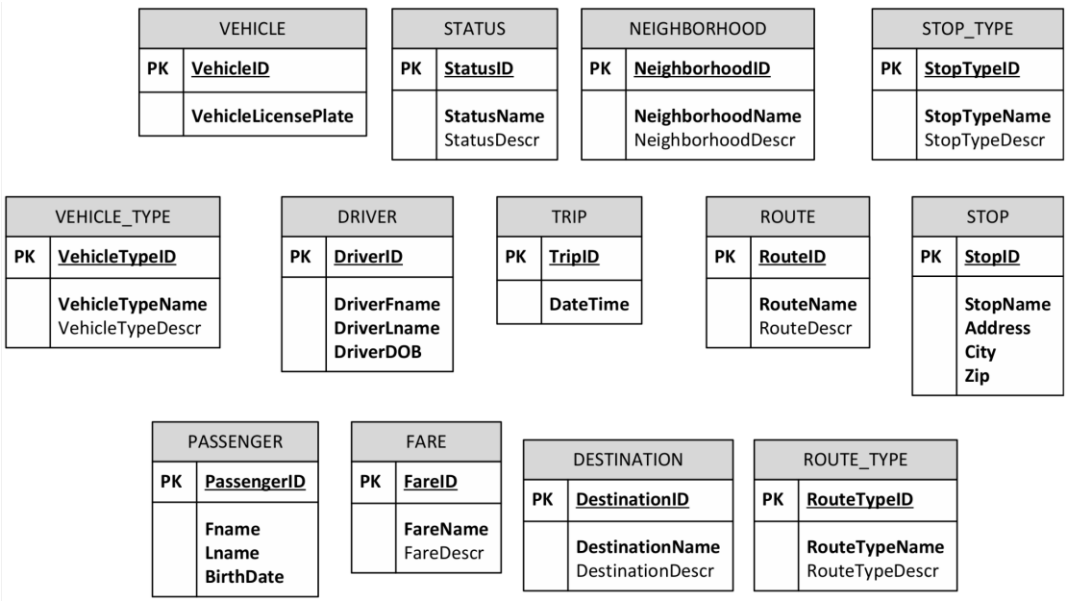


Figure 4.16: Moving entities into common sections based on possible relationships

Anytime we are designing a database, it comes together more easily if the related entities are clustered nearby. While not required for functionality or performance, it helps the design process 'make sense' when relationships are ultimately defined with shorter lines that do not cross over each other. These are more elegant and tremendously easier to read!

Relationships will change!

As the design of any new database is being fleshed out, there will often be situations where the definition of a relationship will change as we learn more about the data and continue through the normalization process. There will be several concrete examples coming up where previously established relationships are modified. Be open and flexible to change!

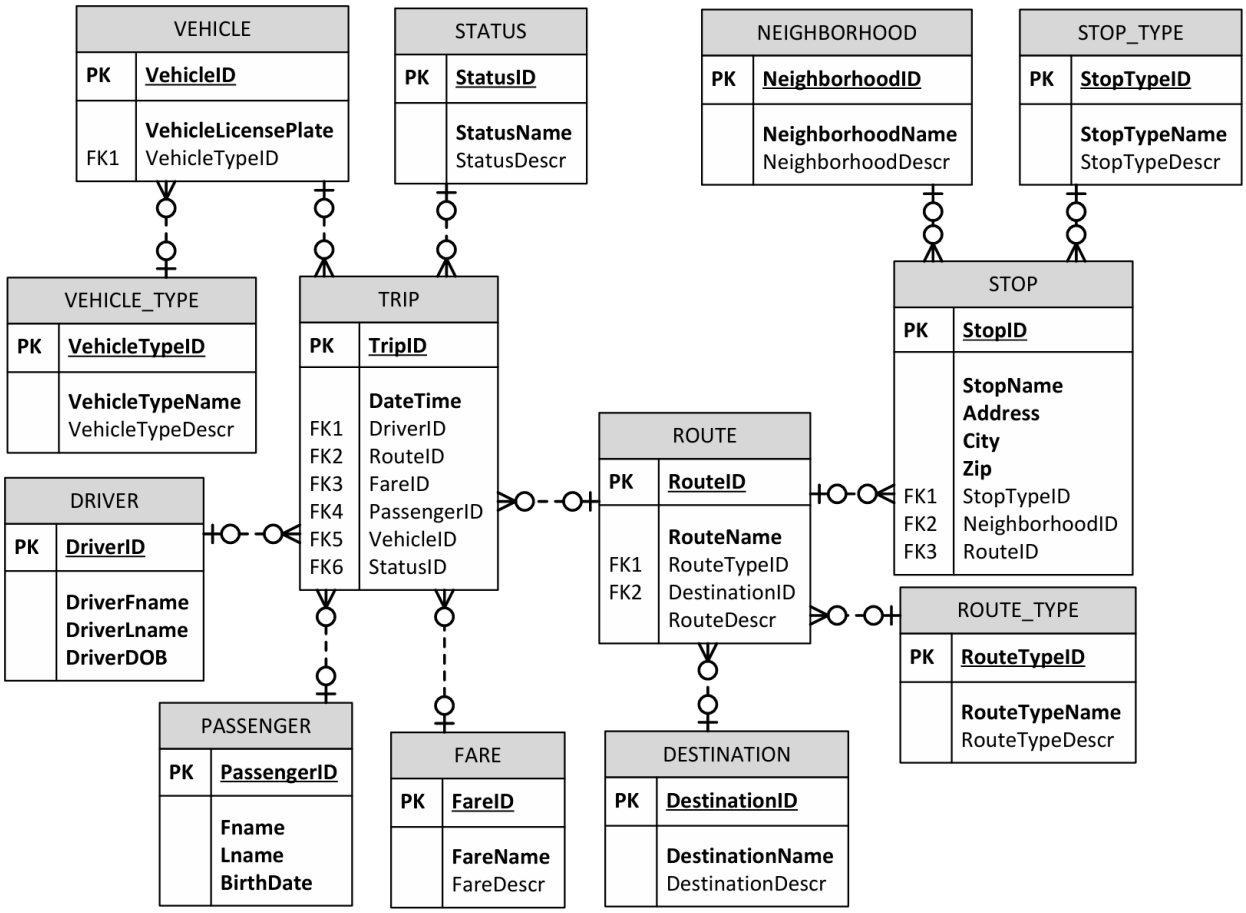


Figure 4.17: Creating initial relationships with PK/FK

As we continue in this ironing process, we are beginning to assemble a collection of entities that will begin to look like the finished ERD in chapter 3 *Data Modeling and Normalization* with Figure 3.24 that we saw previously. Let us analyze what we have so far in Figure 4.17.

The entities at this stage of development are going to look different than the entities initially created during the 'vertical' method as seen in Figure 3.24. This ironing/horizontal method is not going to be as exhaustive as the traditional or 'vertical' method where we conducted the brainstorming of all the attributes of each entity in a single effort. The goal here is to follow a template of 'ID + Name + Descr' as basic attributes and quickly move on to the next object effectively creating 'puzzle pieces' that will be aligned into relationships in a still under-developed form.

As entities are added, we want to be aware if they will be holding 'transactional' data or whether they will be providing context to the transactional entities. In simple terms, an entity that has foreign keys will be collecting transactions while an entity that has no foreign keys will be providing context. We may call these 'look-up' entities. Most databases will have more look-up entities than transactional.

As was stated in earlier chapters, designing relational databases requires us to think in terms of designing for machines where each object is isolated and 'tagged' with labels as opposed to designing for people, where the data needs to make sense when we read it or where it is stored. An example of designing for people is a garage sale, where a group of items like 'kitchen goods' or 'electronics' are stored in a common box or section of the garage. With relational database design we must break that line of thinking! We are tagging items with their type as opposed to lumping all a particular type into a common entity.

Look at both DRIVER and PASSENGER tables; two entities that are the people (arguably the most important aspect of any business). There are only factual data elements about each person held in the tables 'about them'. No elements about what they did, where they did, or when they did it should be in their base table. The transactions that tell the story of their behavior or actions will be recorded elsewhere.

Data is not stored on disk in any 'shape' or cluster for people to 'see as it is stored'; we will always be using the SQL language to re-assemble data in a precise form to answer questions; these are very different from storing data in the shape of the query on disk!

This idea of having data not making sense for people as it is stored is perhaps the most difficult modification of thinking for new designers.

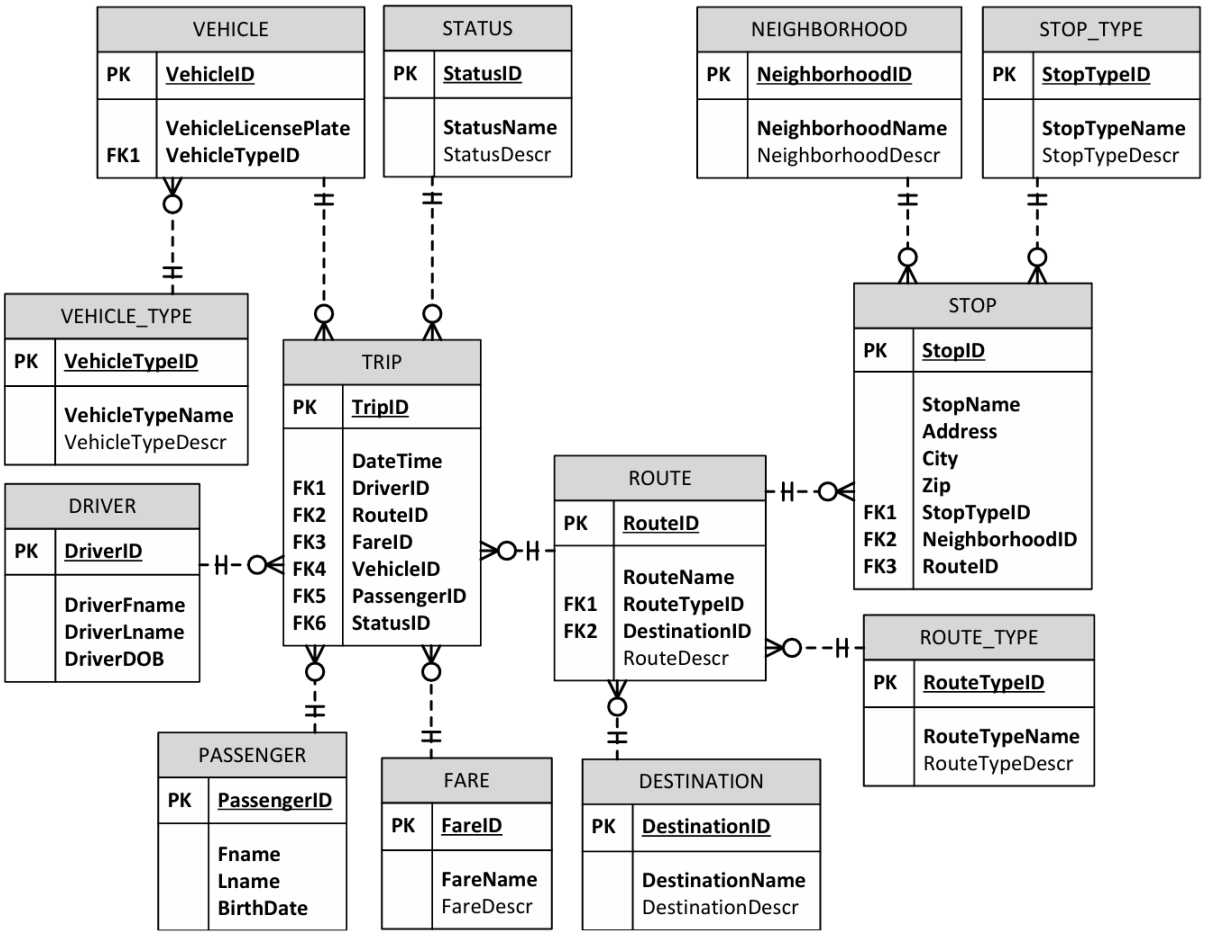
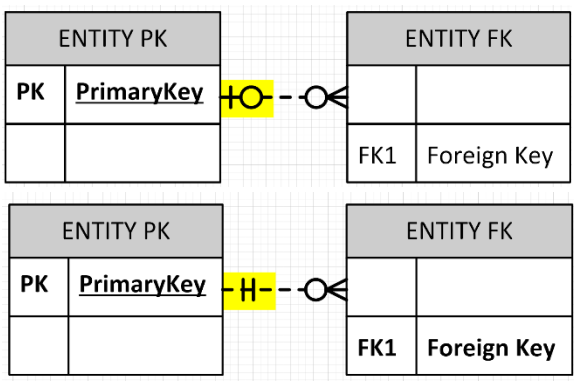


Figure 4.18: Making FK values and any other attribute mandatory where appropriate

Now the initial relationships have been added with primary keys and foreign keys. We will need to take a close look to determine if the foreign keys can be NULL; many designers argue against allowing NULL (empty) values if possible, but each database is going to have distinct choices to be made on this. When the decision is made to make some or all the foreign keys mandatory, the modeling tool will make them bold as well as changing the crows' foot notation from an 'O' to a 'dash'. Please refer to an example found in Figure 3.44 from chapter 3 *Data Modeling and Normalization* as a reminder (shown again below):



Review of Figure 3.44: Example relationships of mandatory and optional

Many designers consider re-reading and evaluating previously defined relationships may initially seem cumbersome or even slightly annoying. This is a natural reaction! Embrace this process instead as an opportunity to gain intimate knowledge of the data model if not the entire mission of the business organization. This is a valuable expertise that is only obtained via thorough observation, analyses, and study.

Next in the process of re-evaluating relationships looking for many-to-many include discovering several are in fact M:M. These include the relationship between TRIP and STATUS, ROUTE and STOP, PASSENGER_TRIP, and finally ROUTE and DESTINATION. In each situation of re-evaluation, we read the relationship in both directions (left-to-right as well as right-to-left).

We will discover that these four relationships are ‘many-to-many’ and will need to be resolved. This means adding a new entity called an ‘associative entity’ in between the affected entities in each relationship to create multiple ‘one-to-many’ relationships. These are highlighted in purple below in Figure 4.19.

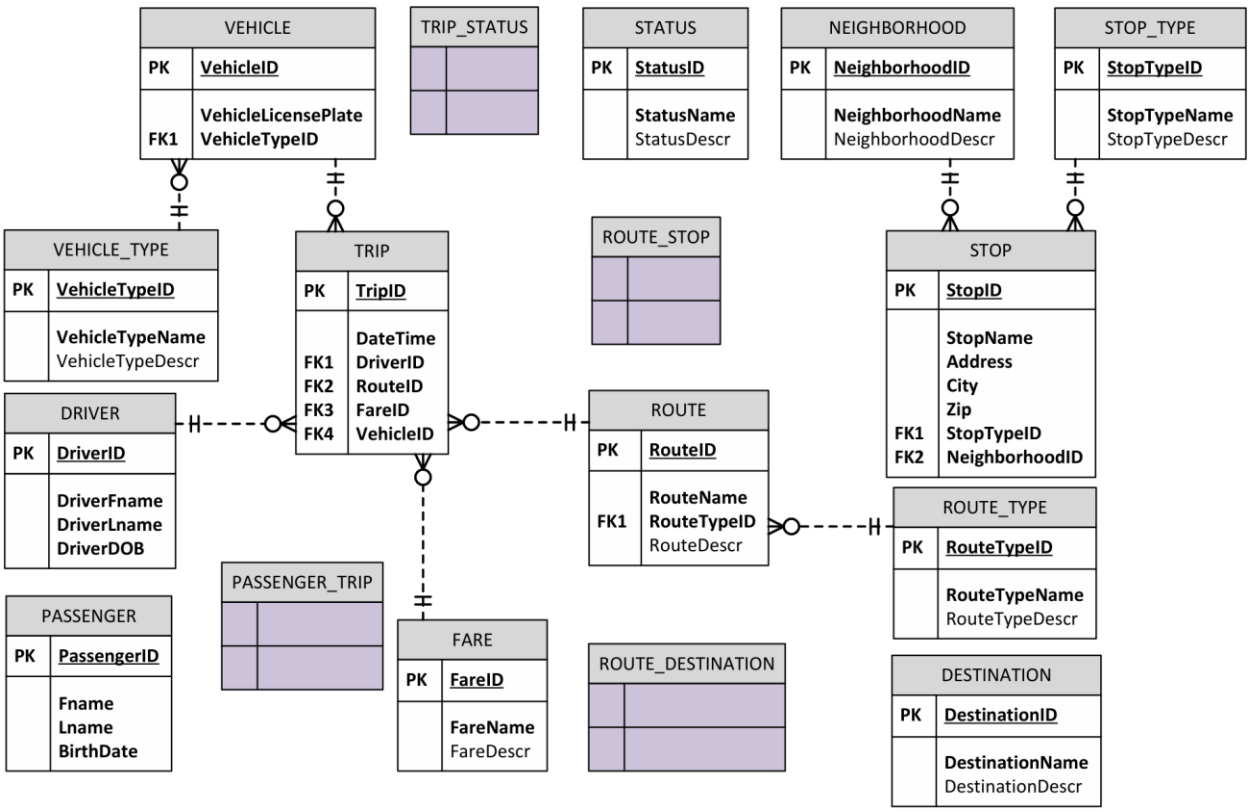


Figure 4.19: Re-Evaluating each relationship looking for M:M

After adding the initial collection of entities, we want to re-evaluate them in the context of the other entities. Our goal is to see if these new entities have changed our perception of any other relationship. Upon further evaluation, it looks like there are more M:M relationships than before:

- PASSENGER : TRIP (M:M)
- TRIP : STATUS (M:M)
- ROUTE : STOP (M:M)
- ROUTE : DESTINATION (M:M)
- TRIP : FARE (possible change…not liking this as a relationship anymore)
- STOP : NEIGHBORHOOD (no change)
- TRIP : ROUTE? (no change)
- TRIP : DRIVER (maybe M:M; Not worth the complexity for once a year issue)

The first four relationships that are re-evaluated render M:M and require resolution with a new entity. Also, after evaluating TRIP : FARE, it is becoming more apparent that a Fare does not describe a TRIP but perhaps something involving passengers.

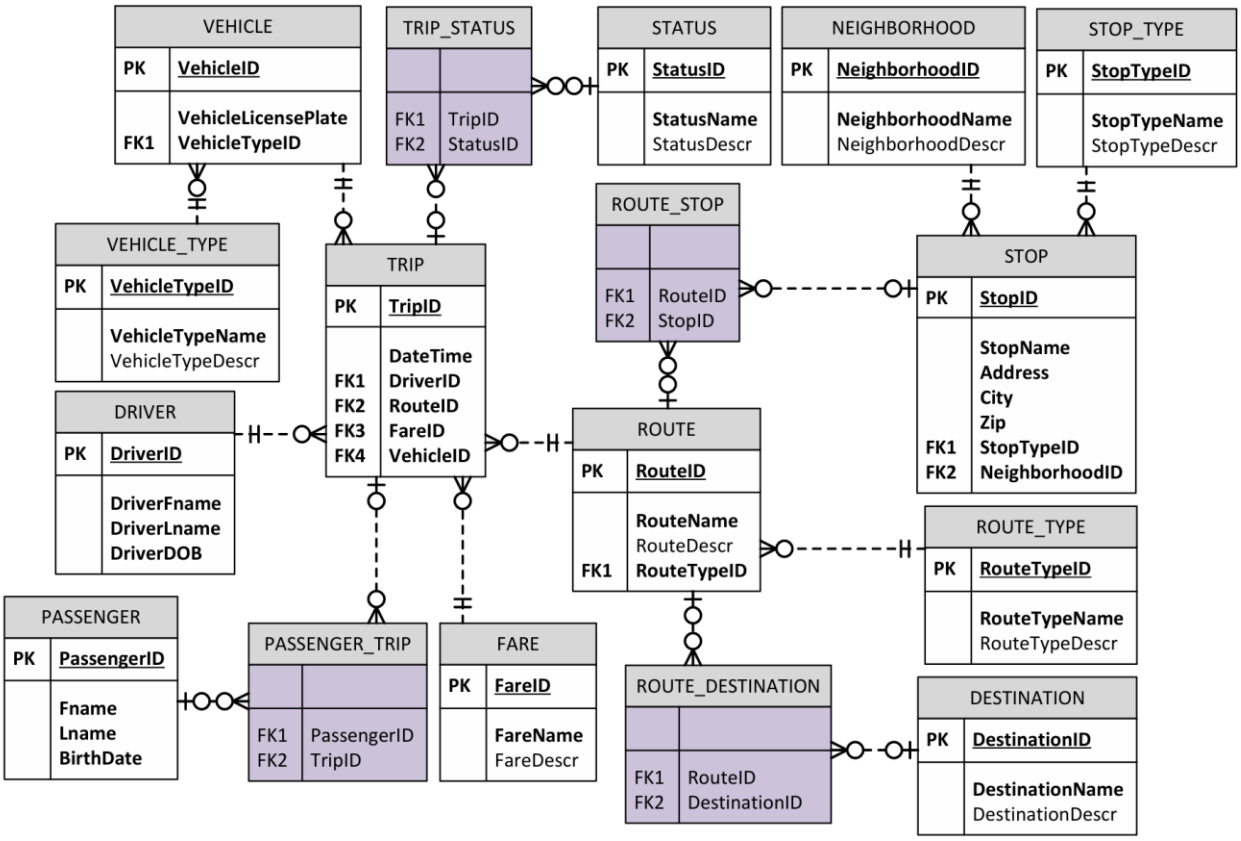


Figure 4.20: Resolving several M:M relationships

The process of taking the new associative entities discovered in Figure 4.20 is to pull inward the primary keys of each participating entity. Our initial option is to have composite primary keys assembled from the foreign keys just brought in, but chances are high that the better choice is to also assign a surrogate PK for these associative entities.

Resolving M:M

Note: a quick reminder of how to resolve a M:M relationship is to place an entity between the affected entities and immediately 'flip the forks'. Another reminder is the forks always (seriously, always) point inward to the new associative entity. The final phrase to remember is 'the foreign key follows the fork'. So, putting these all together, we know that a M:M requires a new entity in the middle, followed by bringing in FKs from the affected entities. See if the next few examples make sense!

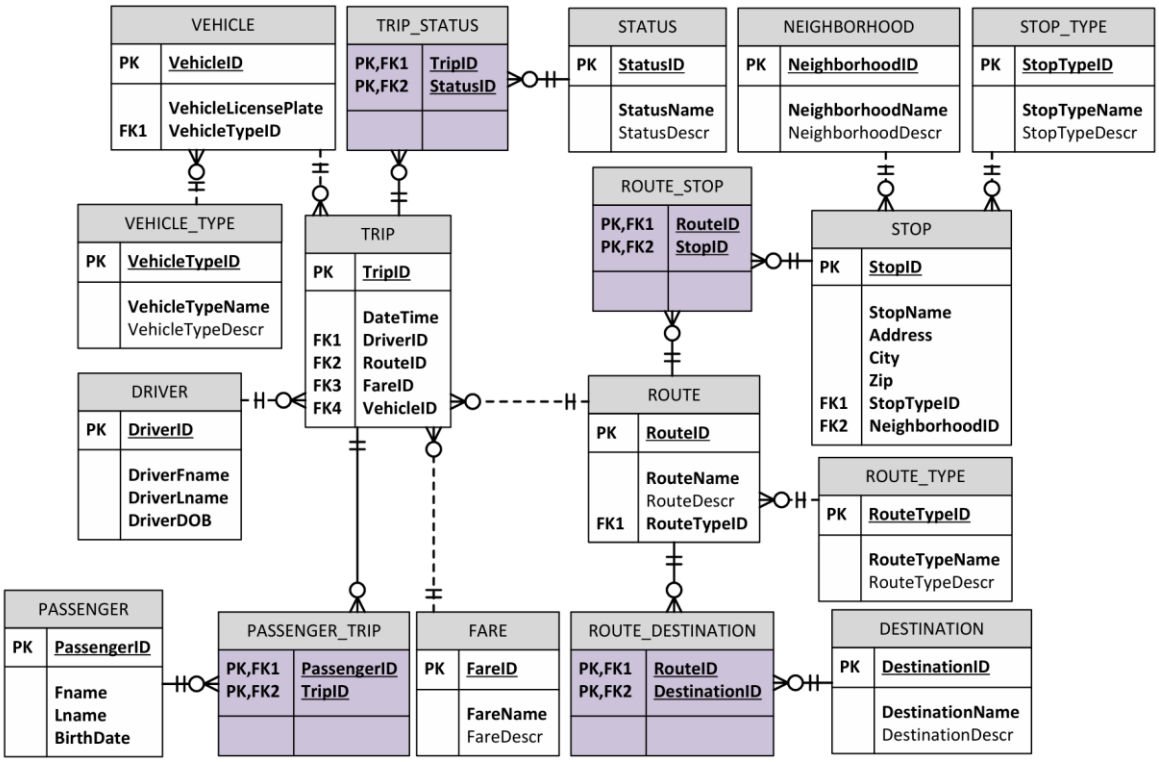


Figure 4.21: Exploring the 1:M primary keys as composite PK

After creating these new entities, we may want to consider a more friendly set of names for the entities. Often the compound nature of entity names (PASSENGER_TRIP for example) is created to simplify the pedantic nature of normalization. Both PASSENGER_TRIP and ROUTE_STOP sound weird (especially to folks from the business and marketing side of the organization). We are considering adding new names to include in the diagram to reflect the vernacular used by our primary stakeholders. Now may be the best time to re-label an entity if there is a more germane name that people can agree on.

For PASSENGER_TRIP a better object name is BOARDING as this is a common term used by non-technical folks throughout the world. Additionally, the term SCHEDULE seems to be more appropriate for ROUTE_STOP.

SCHEDULE represents a prescribed calendar of all scheduled routes and when each is supposed to arrive at a specific stop. Encompassing the entire transportation network for the city, representing all scheduled stops across all routes, neighborhoods and stops. As such, there may be a million rows or more (quick math on roughly 50 routes, 12 stops each route, and 25 trips per day for 3 months before the next revision of schedule). This table is

also relatively static as in 'read-only' with very few if any updates between revisions. It may only be updated 3 or 4 times each year when the entire system is refreshed with slight seasonal changes.

Be careful! This is getting a bit tricky again and there may be a need to change the definitions of which entities are connected to others. Anytime changes are made to entities or relationships, like resolving M:M with associative entities, we will need to review the design quickly to ensure other relationships have not been affected as well.

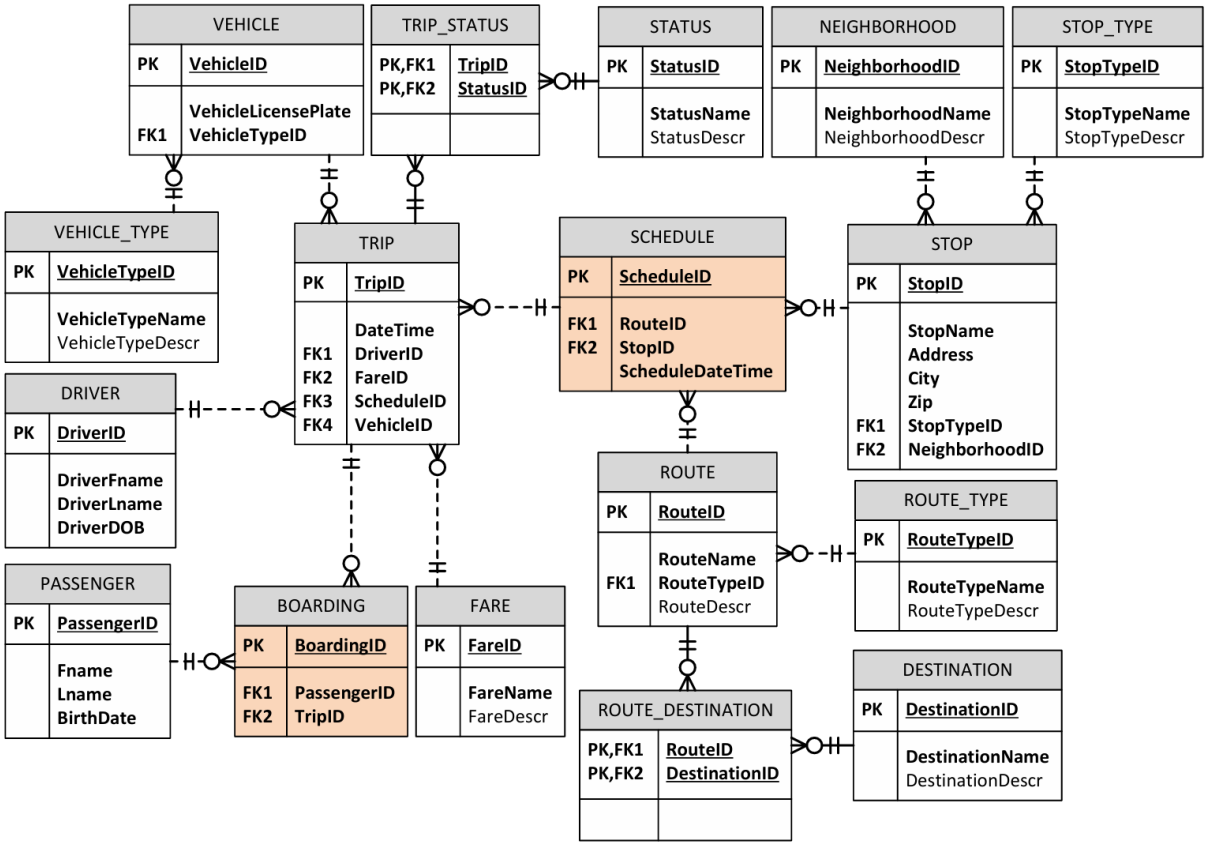


Figure 4.22: Renamed entities SCHEDULE and BOARDING with surrogate Primary Keys

As can be seen in Figure 4.22, the entities highlighted in purple have been re-named as well as having surrogate primary keys created. SCHEDULE has also had a new column added to help answer questions about when a particular trip is supposed to be at a specific stop.

In METRO_TRANSIT, it makes sense to re-evaluate what an individual occurrence of a SCHEDULE and a BOARDING represent. It is important to make sure the database is capturing the events in a way that aligns with what the business expects and that reports generated from the database can answer very specific questions.

As we are re-evaluating the design, we should take time to answer the following questions:

- Does a boarding need to include the StopID for each passenger entering the vehicle?
- Does it benefit the design to include the actual time a passenger boards a vehicle?

Let's say it has been determined TRIP needs more information about time and location than was available with ROUTE. We are now changing the database design! TRIP has had the relationship to eliminate ROUTE and instead it will be connected to SCHEDULE.

Please consider the updates as follows in Figure 4.23 that have several changes:

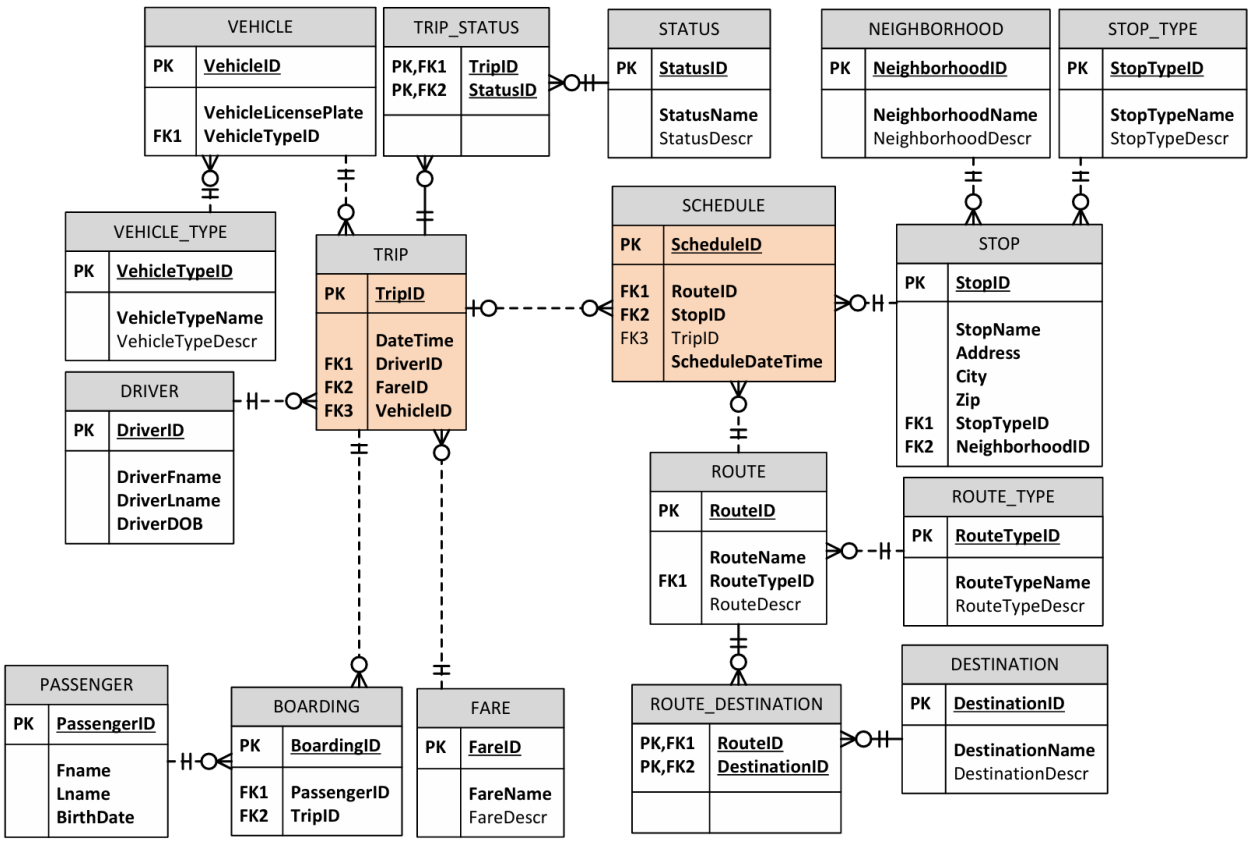


Figure 4.23: Modifying the relationship between TRIP, ROUTE, and SCHEDULE

As seen in Figure 4.23, the revised structure of the ERD has replaced the relationship between ROUTE and TRIP to be now SCHEDULE and TRIP. Please note the direction of the 'fork' in the new relationship is now going toward SCHEDULE. The only slight wrinkle is TripID as a foreign key is SCHEDULE needs to be empty ('NULL') when rows are first added to the table. This is important!

We must slow down and consider what knowledge of data people in the system will have and when that knowledge will be available. In the relationship between TRIP and SCHEDULE,where a schedule is prepared and published to the world three months ahead of time, we will not be able to know for sure which driver or exact vehicle will be on shift 3 months ahead of time (at least not with significant accuracy).

By allowing the system to allow empty values for the exact TripID, we allow for flexibility in being able to communicate to passengers who need to plan excursions while also being able to update the system when a TripID is available. Since a single trip has a dozen or so stops and the potential of more than 100 passengers, it will benefit the business to record which stops have the most boardings. Previously there was no way to know for sure where a boarding occurred.

As we evaluate the effects of the changes made in Figure 4.23, we want to address how it affects the entity BOARDING. Originally, it made sense to align a passenger to a trip; after the developments and changes made along the way, consider changing the definition of a boarding to be connected to SCHEDULE.

Now that BOARDING has been created, it appears to be a great time to change FARE to be attached to BOARDING. This change (which includes adding more columns such as FareAmount, BeginDate, and EndDate) allows the business to notify the public when fares increase, in addition to tracking different fares per boarding.

Consider the next rendering of METRO_TRANSIT, which has several changes based on re-evaluating the entities and relationships highlighted in color:

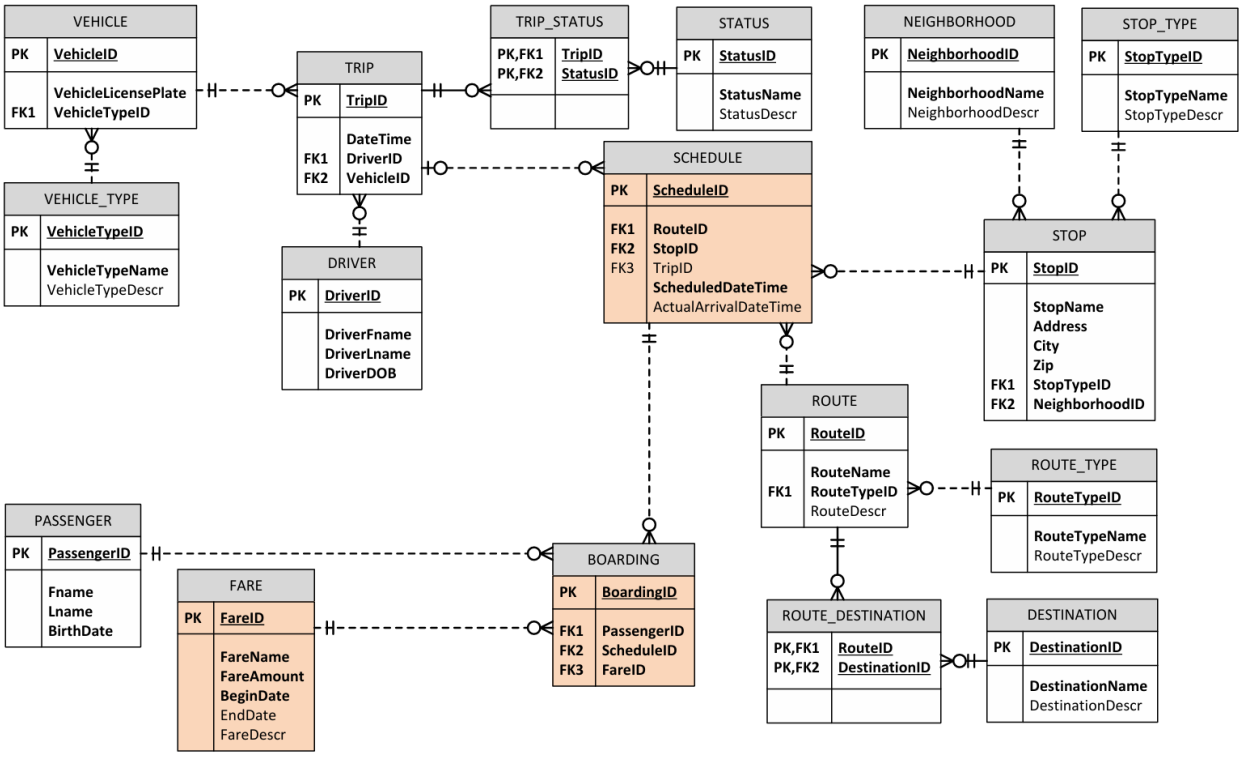


Figure 4.24: Re-evaluating FARE, BOARDING, and SCHEDULE

The changes seen in Figure 4.24 reflect an updated and evolved interpretation of what a BOARDING is as well as the exact data that may be desired to answer future questions regarding a SCHEDULE. These updates require changing which entities are connected to each other! In this new rendering, BOARDING no longer has a relationship with TRIP and instead relates to SCHEDULE. At first glance, this may be slightly confusing or simply unclear; that is normal. The decision to make these updates is based on getting more data that informs each BOARDING. Consider the following:

- SCHEDULE has FK values RouteID, StopID, and TripID
- Additional columns include both the ScheduledDateTime and ActualDateTime

Since SCHEDULE contains these details, significantly more questions can be answered about each BOARDING, including where each occurred on a Trip, as well as which were

before/after scheduled pick-up times. Further evaluation and ironing of the design should include exploring if adding 'type' entities onto base objects will benefit the design. We should think in terms of possible categories of the base objects/entities that the business may reference. Does the design benefit from having a type of fare, vehicle or passenger?

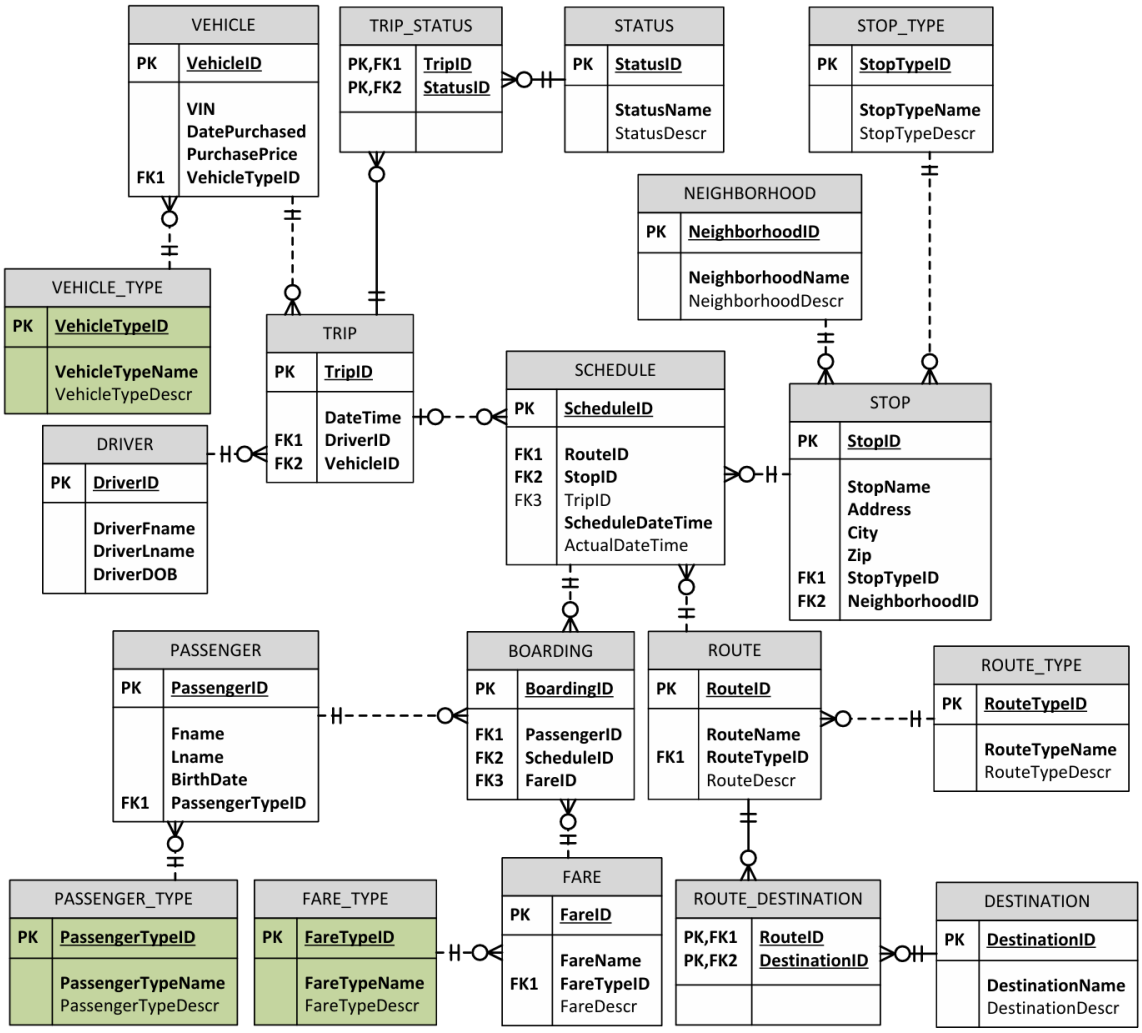


Figure 4.25: Adding several TYPE tables to existing entities

This is the part of this method of database design where we begin to explore more attributes as we did more extensively during the more traditional academic process in the 'vertical' design method. In this step of the design, we most likely will review the possible choices with the business side of the house and other stakeholders to validate these additional entities are necessary. Other entities may be discovered because of check-in discussions with stakeholders!

See Figure 4.26 for potential new entities not included in original requirements:

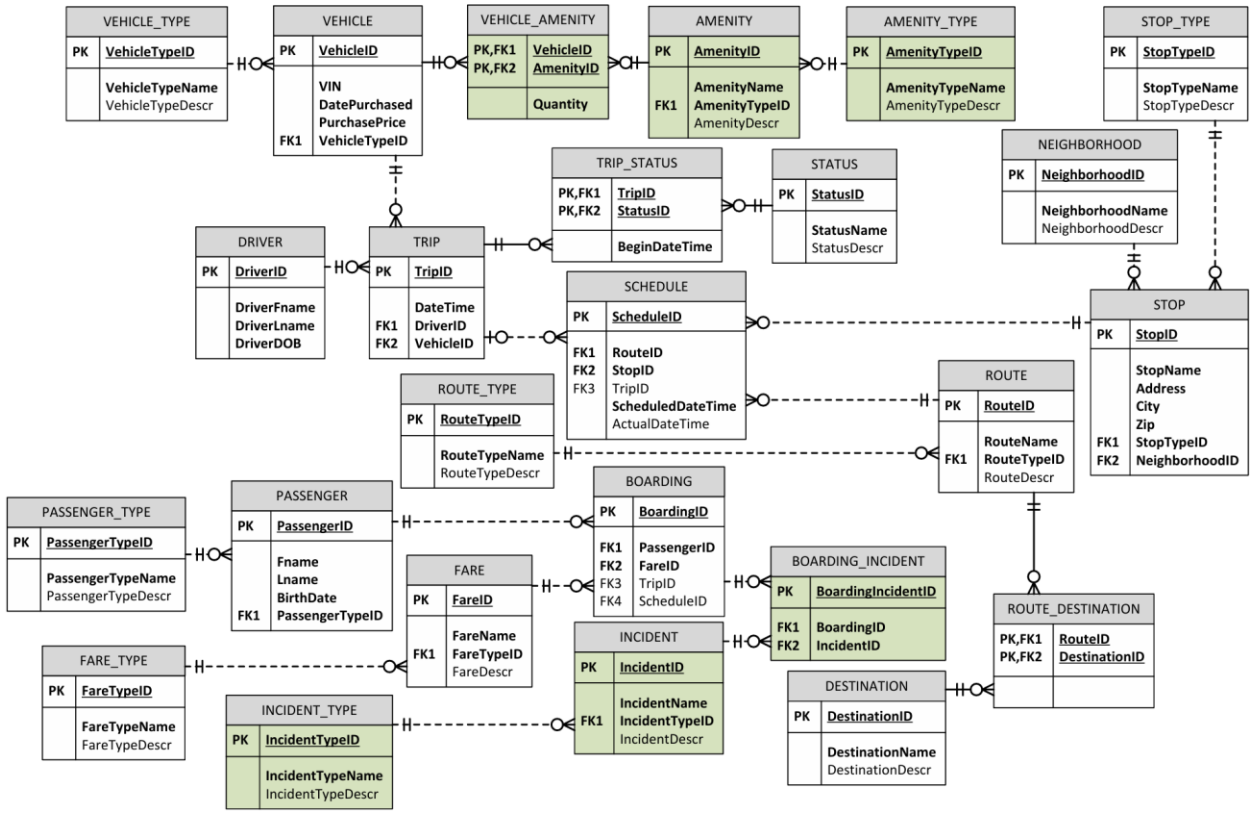


Figure 4.26: Adding additional entities based on conversations with stakeholders

After making these changes, a check-in of sorts with our primary stakeholders is in order only to confirm our thinking and get team validation of our progress. It can be assumed there will be at least one request and/or discrepancy discovered during our check-in with our stakeholders. It is determined that a single BOARDING can be involved in more than a single INCIDENT as well as the same INCIDENT involving more than one person (which translates to more than one BOARDING). This relationship is now M:M and is updated as such in the design.

Highlighted in green in Figure 4.27 are the last few entities requested by stakeholders after a collection of check-ins that reviewed the design and sought feedback on possible improvements to reporting. These include an expansion of the data captured with EMPLOYEE, POSITION, JOB, and SEX. One final entity COMMENT allows for unlimited tracking of the details of any recorded INCIDENT.

Several computed columns are highlighted in yellow. These provide aggregations for important calculations used in weekly reporting perhaps for a dashboard.

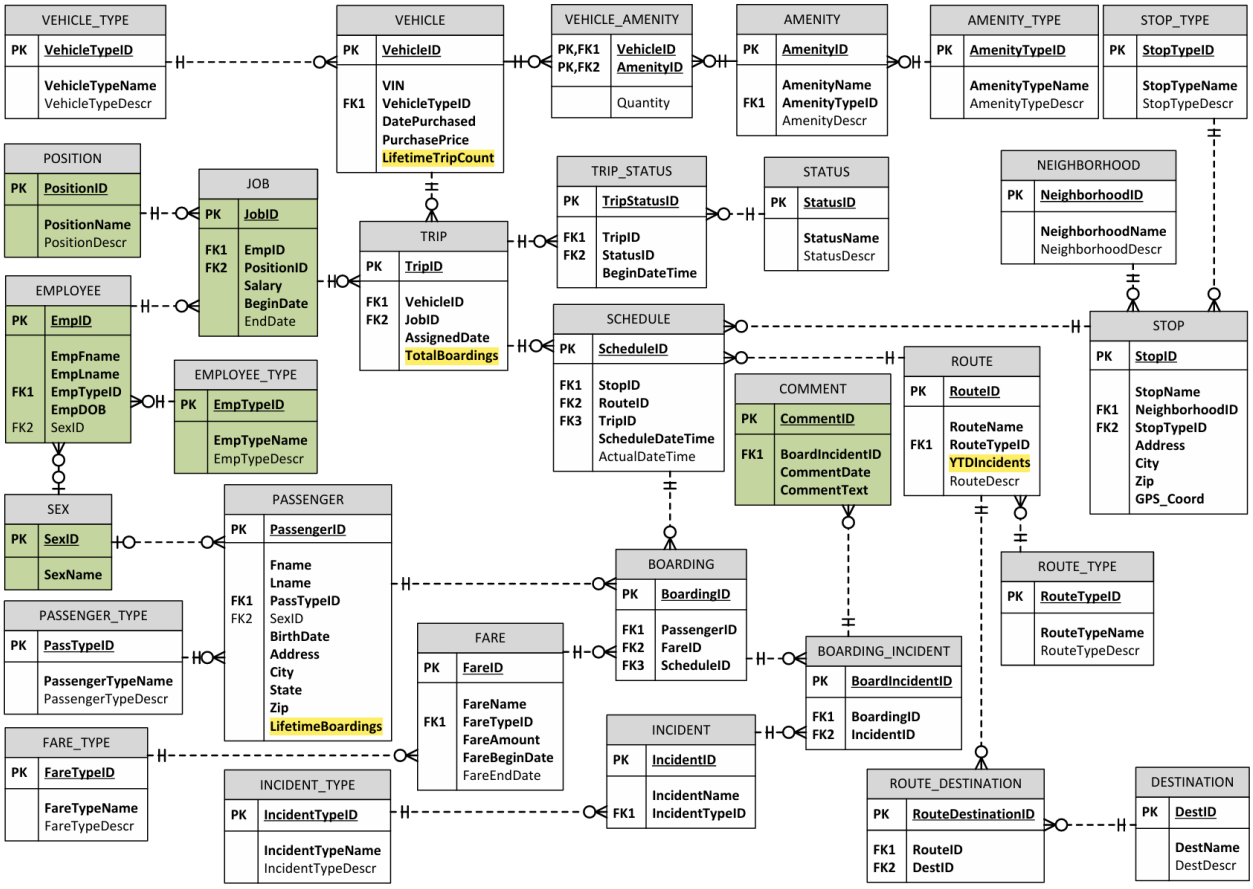


Figure 4.27: Final few entities and reporting columns added to ERD

The diagram in 4.27 is arguably in Fourth Normal Form and is ready to be passed along for having a schema created for deployment.

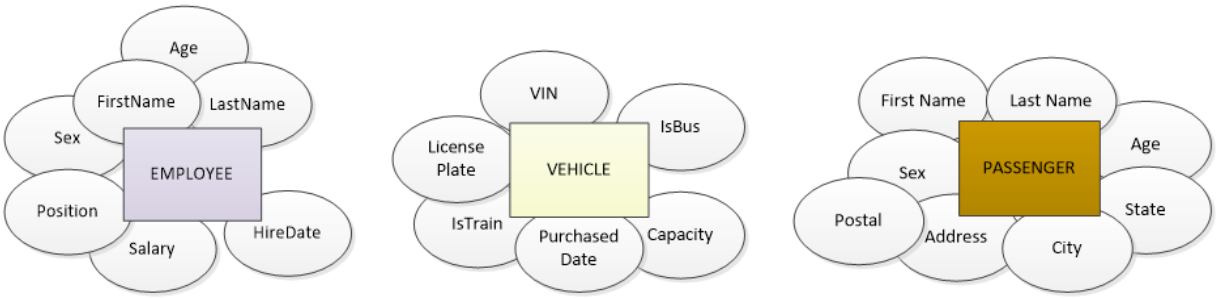
Common Mistakes

When conducting normalization, beginner data modelers often make several of the following mistakes:

- Overloaded base entities
- PK/FK going in the wrong direction (often with a TYPE table)
- Referential loop in the design

When building out a database, it is important to recognize higher-level entities that describe each row of a lower-level entity. These higher-level entities can be thought of as a "category" or "type" entity.

In chapter 3, *Data Modeling and Normalization*, we observed an example of an entity in Figure 3.5 that contained a list of products. A reminder is presented below:



Review of Figure 3.5 example of 'overloaded' entities

It is evident that the VEHICLE entity in this diagram has too many columns or at least is overburdened trying to describe the kind of vehicle a single row represents. While this is important, it is an effective way to manage the data consistently and therefore allows for mistakes. Unmanaged data allows for typos and abbreviations. This is a result of a designer not seeing or exploring a higher-level category entity to take the burden off the base entity. Additionally, there are a couple of "yes/no" questions that appear as column headers (IsBus as well as IsTrain). These columns introduce chaos into the database design!

Also, we observed the range of values allowed under this poor design—there are unmatching values due to misspellings as well as uncontrolled checkmarks of various forms attempting to answer the yes/no questions. To better understand why this design is considered flawed, try to write a query that will account for every possible misspelling, typo, or variation of checkmark from data presented in Figure 3.3 with only 10 rows—imagine trying to prove a query drawing data from a table with millions of rows!

These poor design choices can be very quickly fixed by identifying the problem columns and creating a VEHICLE_TYPE entity to manage the categories. Once VEHICLE_TYPE has been established, there is a finite range of categories that each row in the VEHICLE entity

must adhere to; no longer are there typos or abbreviations that will create havoc in even the most basic query.

We tend to avoid overburdened base entities when following the horizontal design method of 'ironing', but we still want to be wary of columns trying to answer questions that should be in a TYPE entity.

A TYPE table will never have a foreign key

Once we have had a breakthrough by establishing a _TYPE entity for each base entity, it is another common mistake to place the FK in the wrong entity. Remember, once we bring in a foreign key, we are forever limiting a row (which includes a primary key, some kind of name, and a description, most likely) to one and only one occurrence ever!

Let's consider STOP_TYPE entity and mention the different types of bus or train stops:

- Covered
- Uncovered
- Elevated
- Tunnel

If we can only come up with four values, then there should never be more than four rows in this entity! This goes for any _TYPE entity as these are finite categories.

Please consider the following **flawed** example of an FK being placed in STOP_TYPE entity:

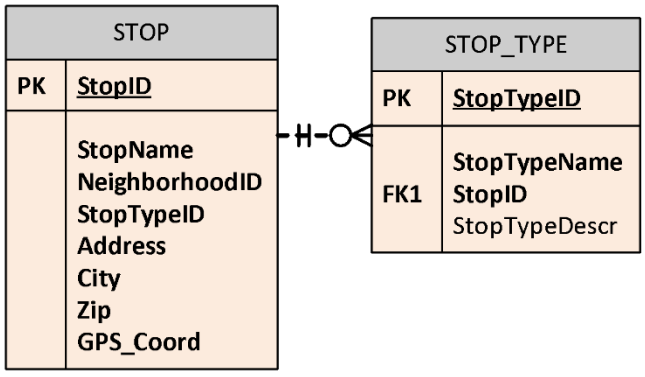


Figure 4.28: Example of a design mistake with a foreign key in a TYPE entity

Please recognize that when we determined values for types of stops, we only came up with four; that means there are only four rows (ever) for STOP_TYPE. When we incorrectly bring in a foreign key, there is a big problem. How are we going to keep track of the tens of thousands of stops throughout a large transit system in a major city like Shanghai or Delhi?

Hmm. We should hopefully see that our first impulse to "connect a base entity to the type" was backward (and following our instinct to lump things in a box like the garage sale example earlier).

A base entity does not describe a "type" (a type does in fact describe a base value). The design in Figure 4.28 would require thousands and thousands of columns in the STOP_TYPE entity to account for the design mistake! One additional column for each stop to align with its 'type'. This is a nightmare!

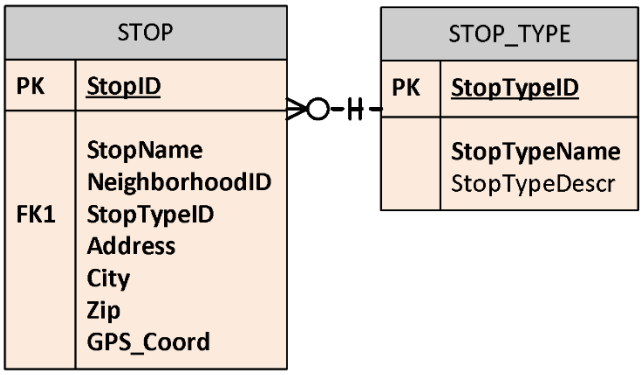


Figure 4.29: Example of a correct design of a TYPE entity providing an FK to base entity

Remembering the rule that a TYPE table 'will never have a foreign key' will go a long way towards recognizing patterns in relational database design and data modeling.

PK/FK going in the wrong direction

Additional common mistakes for beginner data modelers are having the primary and foreign keys "going in the wrong direction". This often occurs due to haste and not recognizing how a foreign key forever limits a row to one and only one value over the

lifetime of the database (or “for a thousand years”, if that is easier to remember). Please observe the following flawed relationship between a PASSENGER and BOARDING:

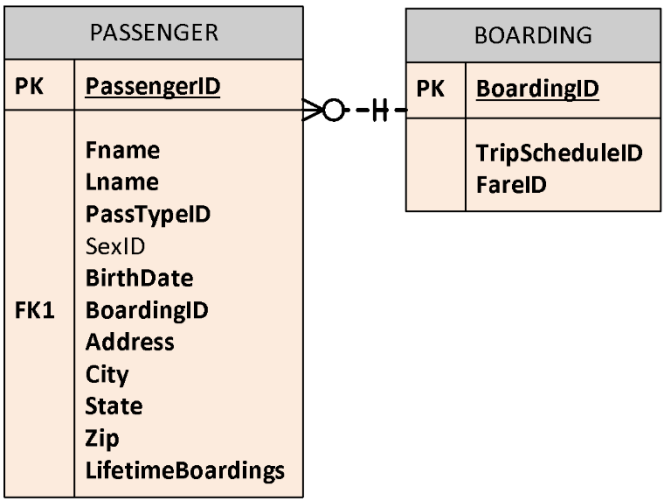


Figure 4.30: Example of a mistake with PK/FK going in wrong direction

Let us slow down and evaluate the incorrect relationship as defined above. We want to track the BOARDING by the specific PASSENGER; we bring in the PK of BOARDING into the PASSENGER entity to answer that concern. On further review, however, we should recognize that bringing in the BoardingID as an FK into the PASSENGER entity forever limits a Passenger to one and only one boarding ever (“for a thousand years”). This ‘1 and only 1 ever’ restriction should set off alarms; every one of us has boarded a bus, car, plane, or train thousands of times. Now we will be limited to just one ever. Wow, that is wrong!

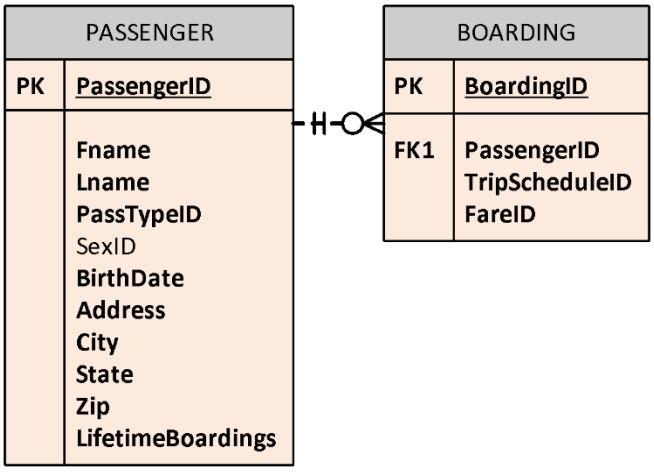


Figure 4.31: Example of a correct design with PK/FK going in right direction

In the above definition, a BOARDING is for one and only one PASSENGER "for a thousand years". Is that true? Yes! Slow down anytime deciding on "where to place a foreign key"; visualize an actual row of data in your head and ask yourself, "Do I forever want to limit a row in this table with one and only one of these FK values for the life of the database?"

Referential loops

A referential loop is when entity relationships form a closed loop. Observe the loop created between SEX, EMPLOYEE, JOB, TRIP, TRIP_SCHEDULE, BOARDING and PASSENGER. The 'LOOP' occurs when users can connect all seven entities in a closed circle by traversing all seven entities in multiple directions.

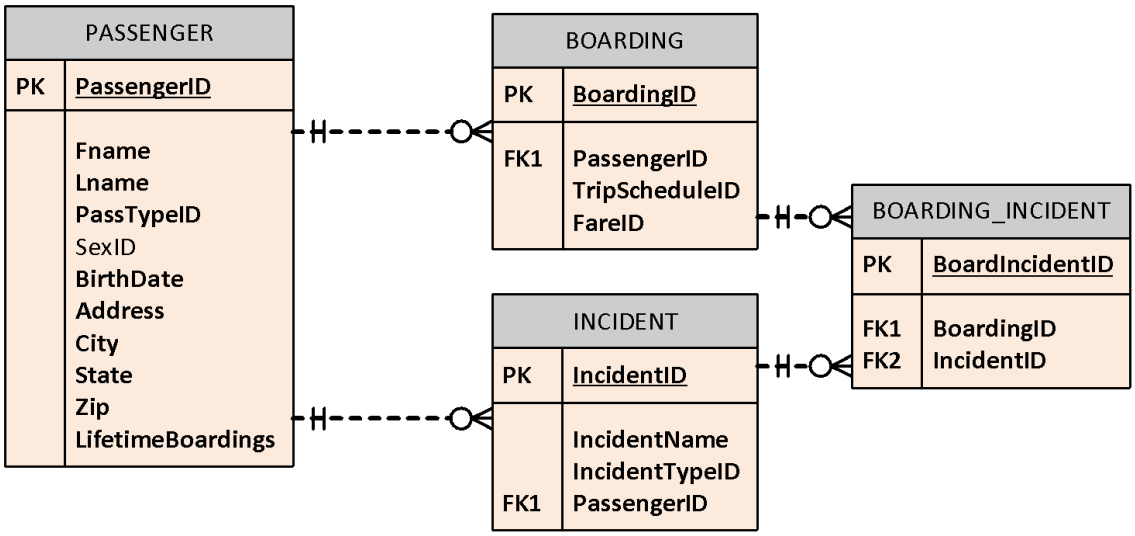


Figure 4.32: Example of a poor design with a referential loop connecting 4 entities

A referential loop places a burden on the user to ensure data remains consistent within the loop. For example, in the above design, PassengerID in INCIDENT must match the PassengerID in BOARDING or there is inconsistent data. Which PassengerID is correct? The best way to resolve a referential loop is to locate the entity that is providing values in two directions or otherwise has both "forks pointing outwards". In Figure 4.32, the entity with 'forks out' is PASSENGER.

We then take a closer look and determine if we can effectively answer questions regarding the INCIDENT (such as which passenger(s) were involved) with a single direction of

relationships as opposed to having the loop. In this example, if we track which BOARDING the INCIDENT occurred, we can assume that the passenger referenced in the BOARDING is the one involved in the INCIDENT. Younger developers often do not trust the database to accurately reference the passenger through BOARDING.

We will learn later in the textbook that this becomes a problem because once a database is being used by thousands of simultaneous connections. With high volume, collisions of users obtaining locks (known as "deadlocks") to make INSERT statements or UPDATES will cause the performance of transaction processing to come to a complete halt. Until then, simply be aware that we want to avoid referential loops if possible.

Please see the corrected version of the referential loop with one of the PassengerID foreign keys removed below in Figure 4.33:

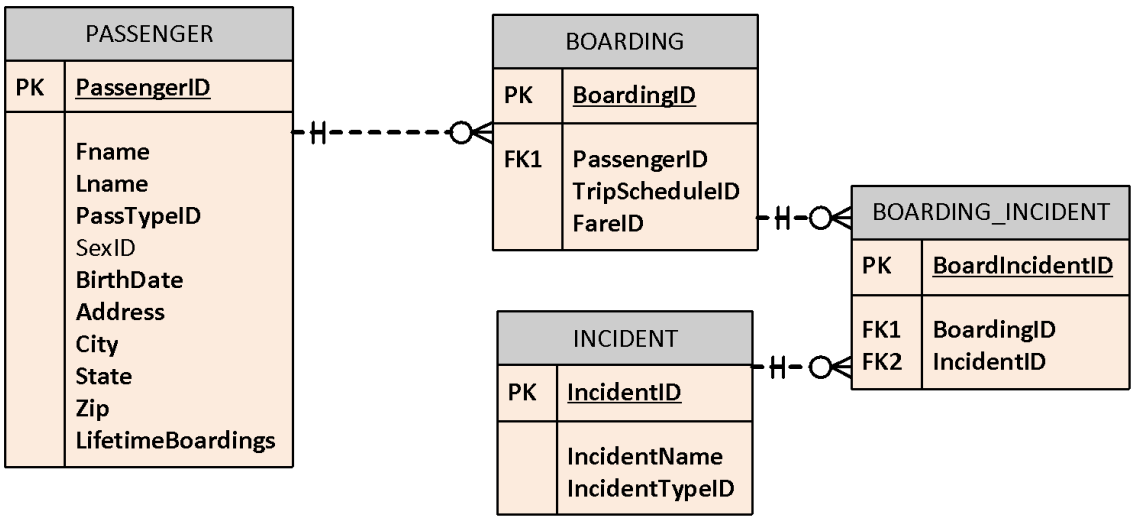


Figure 4.33: Example of a proper design without a referential loop

Again, designing relational databases and normalization are far from easy. Recognizing design patterns will go a long way towards feeling more comfortable and confident in developing robust and scalable database solutions that many people can use. Good work!

Post-Chapter Challenges

Based on your objectives and intentions on learnings from this book, please approach the following challenges as appropriate.

Track 1 (Think): Data Tourist Seeking Ancillary Awareness

Please spend a solid 10 minutes reviewing the following questions, exploring your thoughts. These questions and your responses cut to the essence of this chapter:

- Assuming you have experience riding a public transit system, explore for a minute how the objects captured in METRO_TRANSIT represent those required to operate a real transit system.

- How does a fully normalized database make for better quality data in a relational database? Which objects from real life transit systems might be missing?

- Which of the two normalization processes presented in this chapter (one being the traditional, academic and 'mathy' calculus method and the other being the 'ironing' method) seems to make the most sense to you? Why do you think this is the case?

- Which of the 6 rules that define a relation can you remember if you were to try to explain them to another learner?

Track 2 (Write): Dedicated Student or Recent Graduate

Based on your completion of this chapter, WRITE several paragraphs in response to each question; explore these as if you are being asked a similar question during a job interview!

- Evaluate the process of normalization as presented in this chapter. If followed correctly, is it effective? Why or why not?

- How impactful is a fully normalized relational database as opposed to a database that does not have a fully normalized design? Where or how will this impact be felt by users of the database?

Track 3 (Build): Full Speed Learner Seeking Job

This track targets readers of this book who want to develop professional skills working with data to launch a career or obtain a more satisfying job. Let's begin with the following:

- Imagine you are starting a business in the space of a hobby or interest you have. This can be almost any hobby, such as playing video games, cooking, traveling abroad or exercising. Go through the entire conceptualization and normalization process to make a database. This will be a forever part of your portfolio of learning!

Conclusion

Data modeling and the normalization process is a refined skill that is critical to building robust relational transactional database systems. A well-defined model not only provides a repository of critical organizational data that is trustworthy as well as readily available and secure. A fully normalized relational database also allows insight into the behavior of core company objects (often defined as 'read' activity) without intruding on the processing of capturing new transactional data (known as 'write' activity).

Ultimately, a database is intended to provide organizational learning that enhances the capabilities and competitive strength of an organization as it evolves. Proper analytics with truthful data allow for high-speed learning with the results of companies becoming nimbler and more responsive to customer and market demands. Their investments in people and technologies become deliberate and intentional with greater long-term impact.

Those people who know how to design these structures are effectively independent and capable of conducting their own systems without relying on others for infrastructure or transactional data.

Next up, we are going to finally get into the Structured Query Language (SQL). This is perhaps the single greatest take-away skill from this book and what most of us are interested in. Get ready to be excited for this journey!