Databases and Advanced Data Techniques

Tarapong Sreenuch

Normalisation Concepts and Practices

Which processing environment has more motivation to avoid modification anomalies?

Modification Anomaly

- Unexpected side effect from a row operation
- Must insert, modify and delete more data than desired
- Caused by excessive redundancies
- Strive for one fact in one place

Big University Database Table

StdNo	StdClass	OfferNo	OffYear	EnrGrade	CourseNo	CrsDesc
S1	JUN	O1	2017	3.5	C1	DB
S1	JUN	O2	2017	3.3	C2	VB
S2	JUN	O3	2018	3.1	C3	OO
S2	JUN	O2	2017	3.4	C2	VB

Anomaly Examples

- To insert a course (C4), must know student and offering
- Update multiple rows to change the description of course C2
- A row deletion can cause inadvertent removal of related entities. Deleting first enrollment row (S1, O1) loses details about O1 and C1.

Recap: Modification Anomaly

- Modification anomaly: unwanted side effect from a row operation
- More rows impacted than anticipated
- Motivation for normalization process to remove excessive redundancies

What is the practical usage of falsifying functional dependencies in sample tables?

Functional Dependency Basics

- Constraint on the possible rows in a table
- Value neutral like FKs and PKs
- Asserted
- Understand business rules

FD Definition

- X (functionally) determines Y or X → Y
- For each X value, there is at most one Y value
- StdNo → StdCity if each StdNo value has at most one StdCity value
- X: left-hand side (LHS) or determinant
- Y: right-hand side (RHS)

Unique Constraint Analogy

- Similar to uniqueness constraint
- Place RHS and LHS in a table by themselves

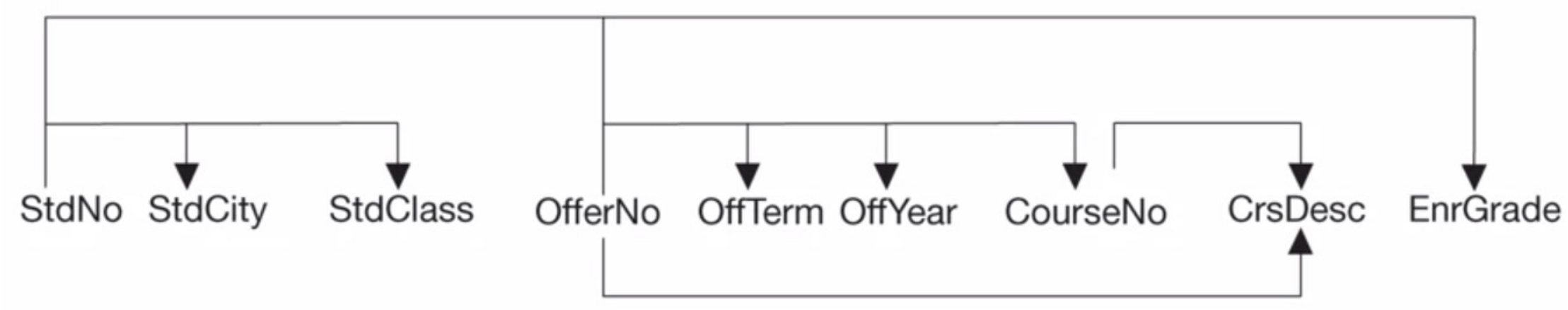
Unique Constraint Analogy

- OfferNo → OffYear
- OfferNo, StdNo → EnrGrade

StdNo	StdClass	OfferNo	OffYear	EnrGrade	CourseNo	CrsDesc
S1	JUN	O1	2017	3.5	C1	DB
S1	JUN	O2	2017	3.3	C2	VB
S2	JUN	O3	2018	3.1	C3	OO
S2	JUN	O2	2017	3.4	C2	VB

FD Lists and Diagrams

- StdNo → StdCity, StdClass
- OfferNo → OffTerm, OffYear, CourseNo, CrsDesc
- CourseNo → CrsDesc
- StdNo, OfferNo → EnrGrade



Falsification of FDs using Sample Rows

- Prove non existence (but not existence) by looking at data
- Help communicate with users about questionable FDs
- Two rows that have the same X value but a different Y value

Falsification of FDs using Sample Rows

- StdNo → EnrGrade

StdNo	StdClass	OfferNo	OffYear	EnrGrade	CourseNo	CrsDesc
S1	JUN	O1	2017	3.5	C1	DB
S1	JUN	O2	2017	3.3	C2	VB
S2	JUN	O3	2018	3.1	C3	OO
S2	JUN	O2	2017	3.4	C2	VB

Recap: Functional Dependencies

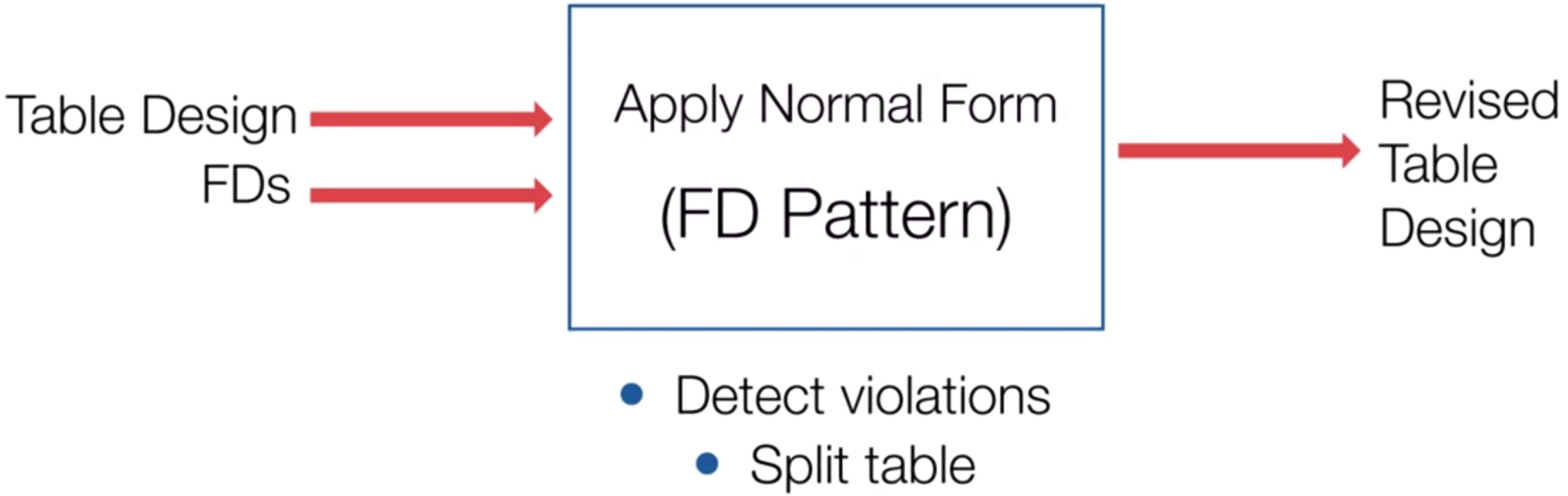
- FDs are important constraints.
- Asserting FDs is essential for removing unwanted redundancy.
- Refinement activity

Who is considered the father of the relational data model?

FDs

Table Design

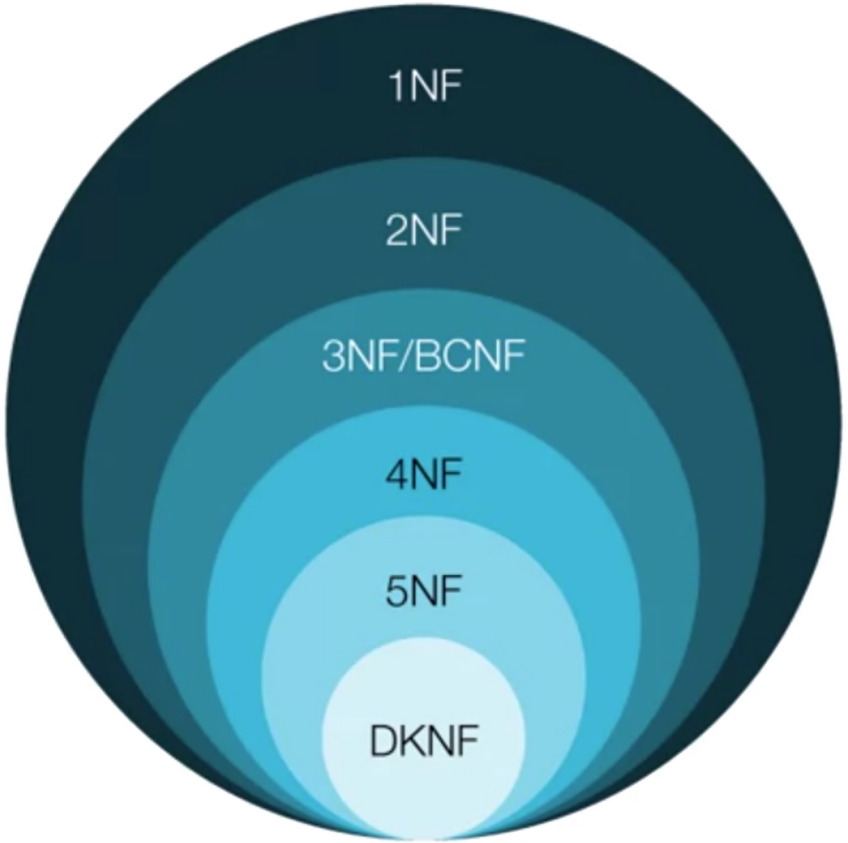
Normalisation



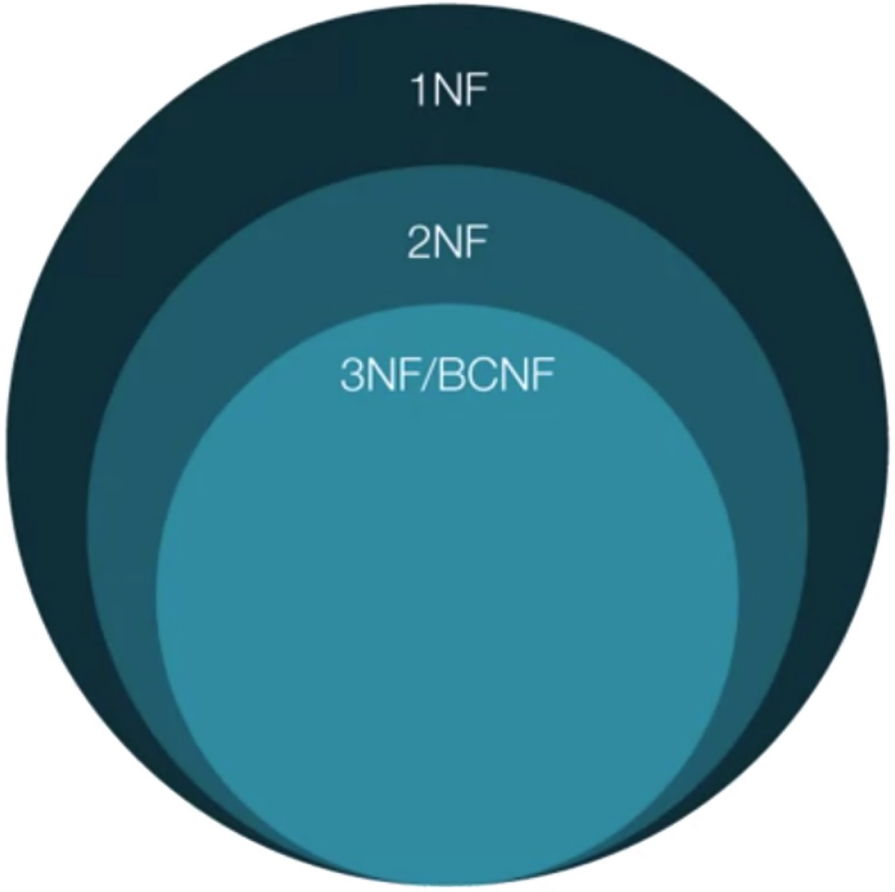
Normalisation Complications

- Determination of complete and minimal list of FDs
- Determination of unique columns from FDs

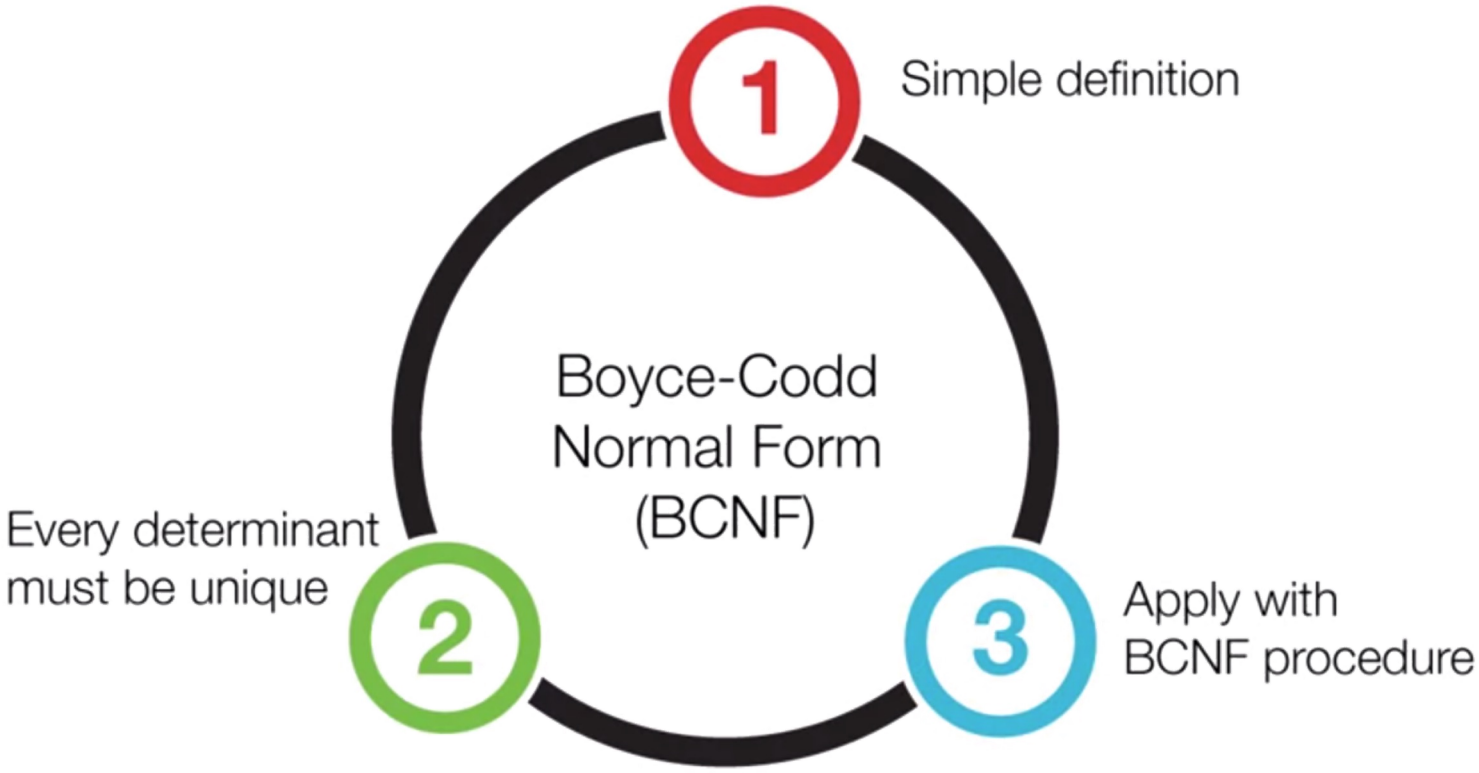
Relationships of Normal Forms



Relationships of Normal Forms



Boyce-Codd Normal Form (BCNF)



Big University Database Table

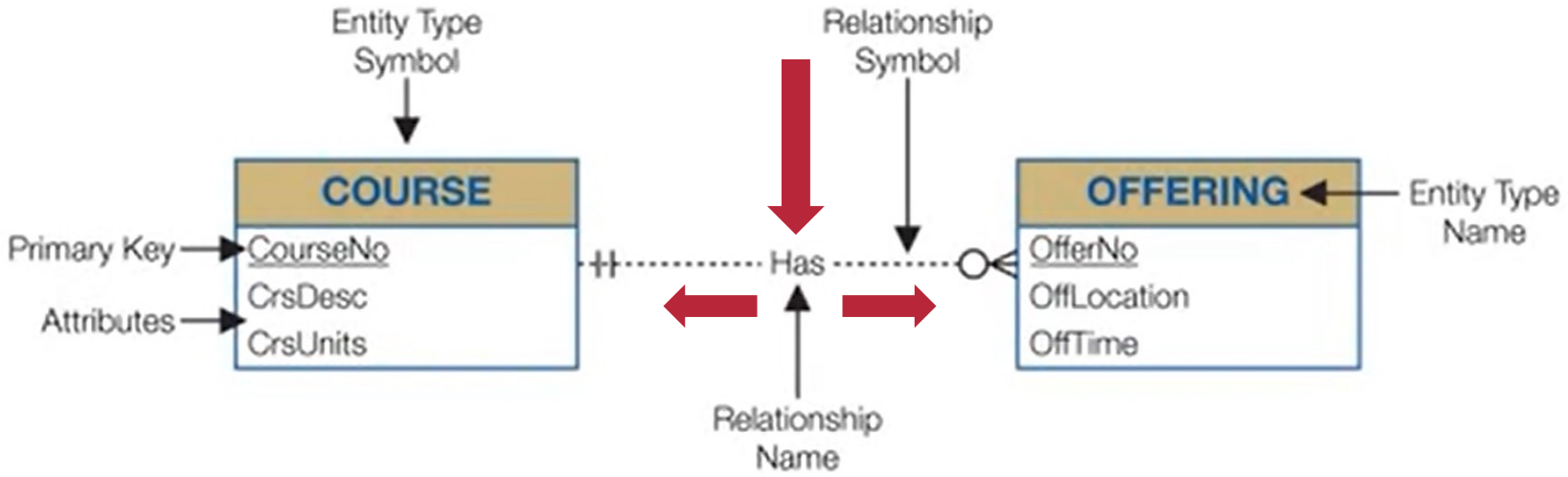
StdNo	StdEmail	StdClass	OfferNo	OffYear	EnrGrade	CourseNo	CrsDesc
S1	joe@bigu.edu	JUN	O1	2017	3.5	C1	DB
S1	sue@bigu.edu	JUN	O2	2017	3.3	C2	VB
S2	mj@bigu.edu	JUN	O3	2018	3.1	C3	OO
S2	tom@bigu.edu	JUN	O2	2017	3.4	C2	VB

BCNF Example

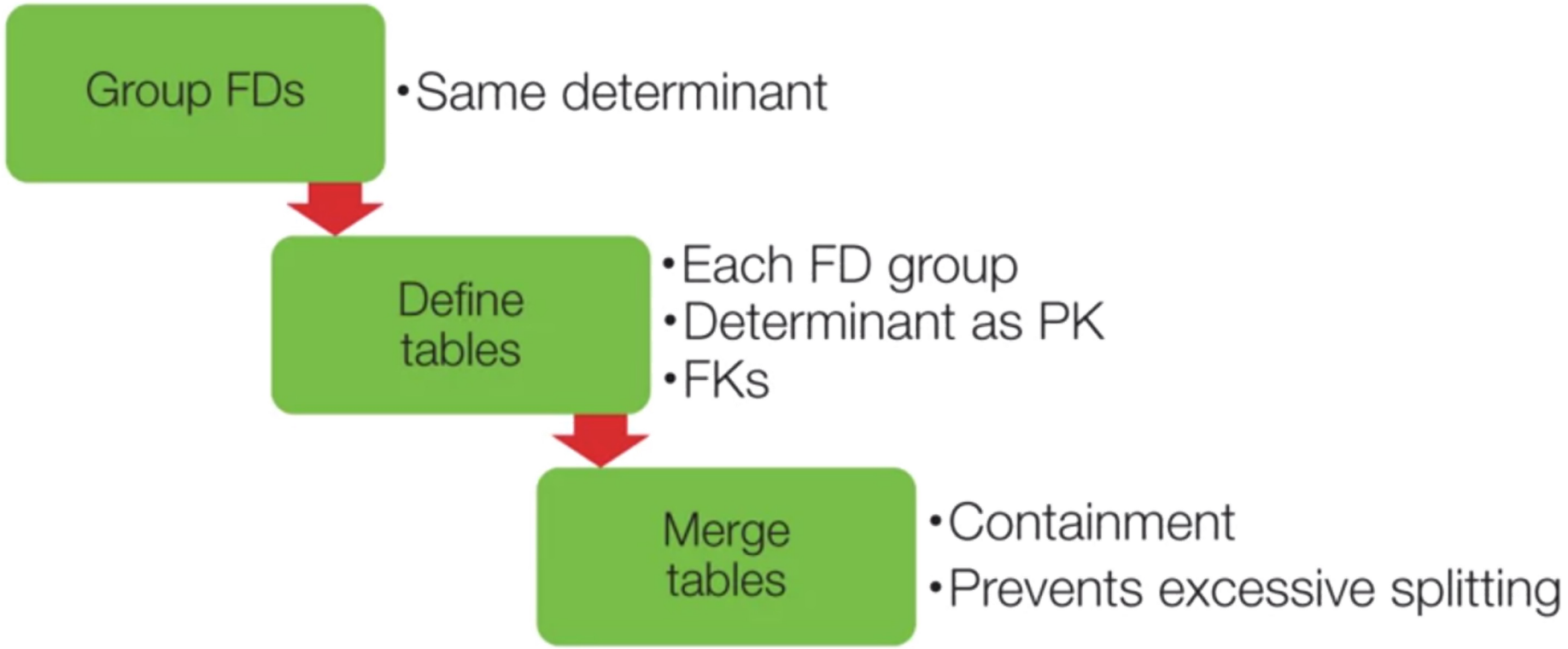
- Unique columns in the Big University table
  - <StdNo, OfferNo>
  - <StdEmail, OfferNo>

- Many BCNF violations
  - StdNo → StdCity, StdClass, StdEmail ❌
  - StdEmail → StdNo ❌
  - OfferNo → OfferTerm, OffYear, CourseNo ❌
  - CourseNo → CrsDesc ❌
  - StdNo, OfferNo → EnrGrade ✅

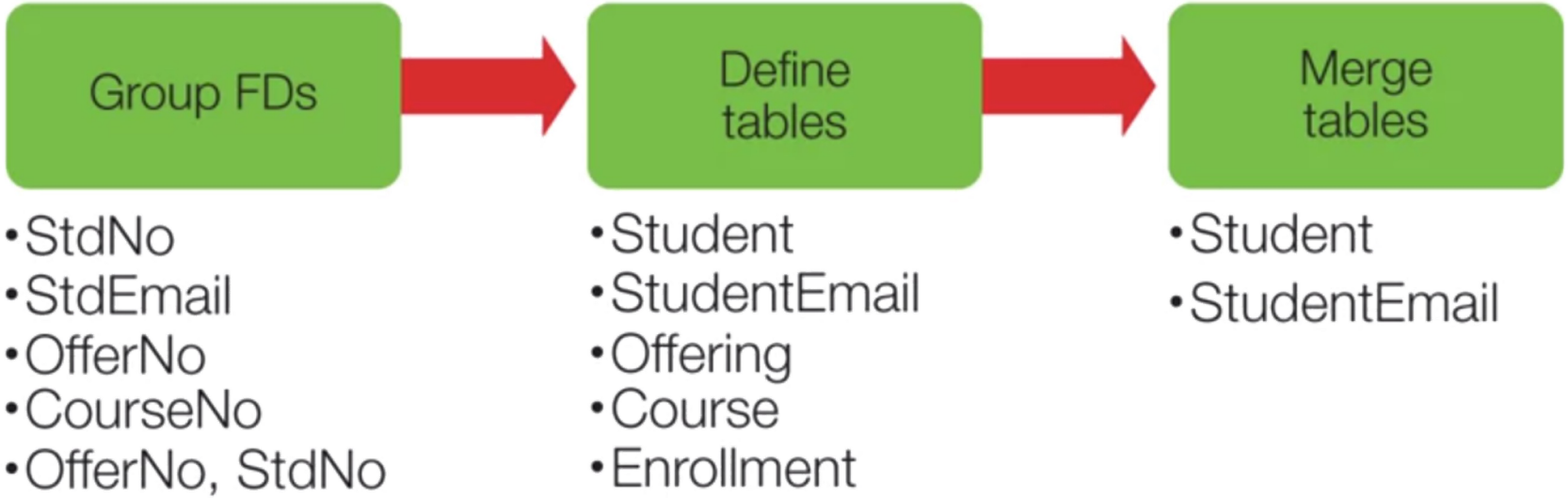
ERD Diagram: Basic Symbols



BCNF Procedure



BCNF Procedure Example



FKs in step 2

- Student.StdEmail,
- StudentEmail.StdNo
- Offering.CourseNo
- Enrollment.StdNo, Enrollment.OfferNo

Merging Tables

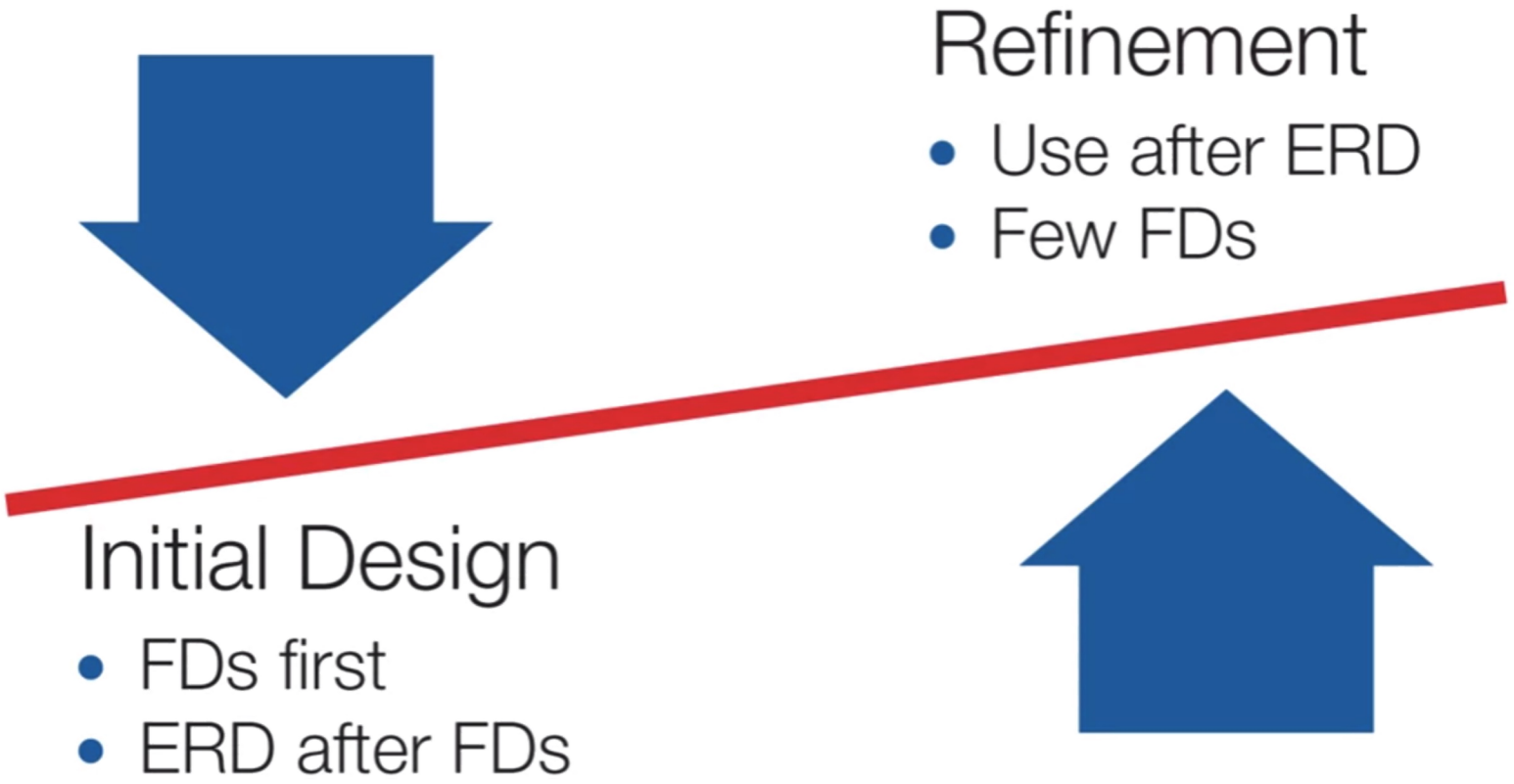
- Step 2 defines too many tables when two columns determine each other.
  - StdNo → StdEmail
  - StdEmail → StdNo

- Merge tables with a containment relationship
  - Student (StdNo, StdEmail, StdCity, StdClass)
  - Student Email (StdEmail, StdNo)
  - Merge tables because Student contains columns of StdEmail

- Multiple unique columns do not violate BCNF

Why is normalisation less important for business intelligence processing?

Competing Roles of Normalisation



Normalisation Importance

- Update biased
- Not a major concern for databases without updates (Data Warehouses)
- Relax normalization sometimes

Denomalisation

- Purposeful violation of a normal form
- Some FDs may not cause anomalies in practice
- May improve performance
- Common for data warehouses

Denomalisation Example

- ZipCode → City, State
- Important for ecommerce business for sales tax
- May be important for ecommerce databases

Recap: Normalisation in Practices

- Use normalization to refine a database design
- Do not lose context of normalisation

Why is the normalisation process simpler but still essential when applying normalisation to a table design after conversion from an ERD?

Modification Anomaly Problem

Big University Table

StdNo	StdCity	StdClass	OfferNo	OfferTerm	OffYear	EnrGrade	CourseNo	CrsDesc
S1	SEATTLE	JUN	O1	FALL	2017	3.5	C1	DB
S1	SEATTLE	JUN	O2	FALL	2017	3.3	C2	VB
S2	BOTHELL	JUN	O3	SPRING	2018	3.1	C3	OO
S2	BOTHELL	JUN	O2	FALL	2017	3.4	C2	VB

Problem Requirements

- Specify one insert, update, and deletion anomaly
- Each anomaly should involve student representation in the table

Modification Anomaly Solution

Big University Table

StdNo	StdCity	StdClass	OfferNo	OfferTerm	OffYear	EnrGrade	CourseNo	CrsDesc
S1	SEATTLE	JUN	O1	FALL	2017	3.5	C1	DB
S1	SEATTLE	JUN	O2	FALL	2017	3.3	C2	VB
S2	BOTHELL	JUN	O3	SPRING	2018	3.1	C3	OO
S2	BOTHELL	JUN	O2	FALL	2017	3.4	C2	VB

Problem Solution

- Insertion anomaly: cannot insert a student (S3) unless an OfferNo is provided
- Update anomaly: must change multiple rows if S1 moves to a different city
- Deletion anomaly: deleting third row also removes details about offering O3 and course C3

FD Falsification Problem

Big University Table

StdNo	StdCity	StdClass	OfferNo	OffTerm	OffYear	EnrGrade	CourseNo	CrsDesc
S1	SEATTLE	JUN	O1	FALL	2017	3.5	C1	DB
S1	SEATTLE	JUN	O2	FALL	2017	3.3	C2	VB
S2	BOTHELL	JUN	O3	SPRING	2018	3.1	C3	OO
S2	BOTHELL	JUN	O2	FALL	2017	3.4	C2	VB

Problem Requirements

- List possible FDs with StdCity as determinant (LHS)
- Identify at least one falsification if it exists for each FD
  - Pair of sample rows for an FD falsification
  - Same LHS (determinant) value in each row but a different RHS value

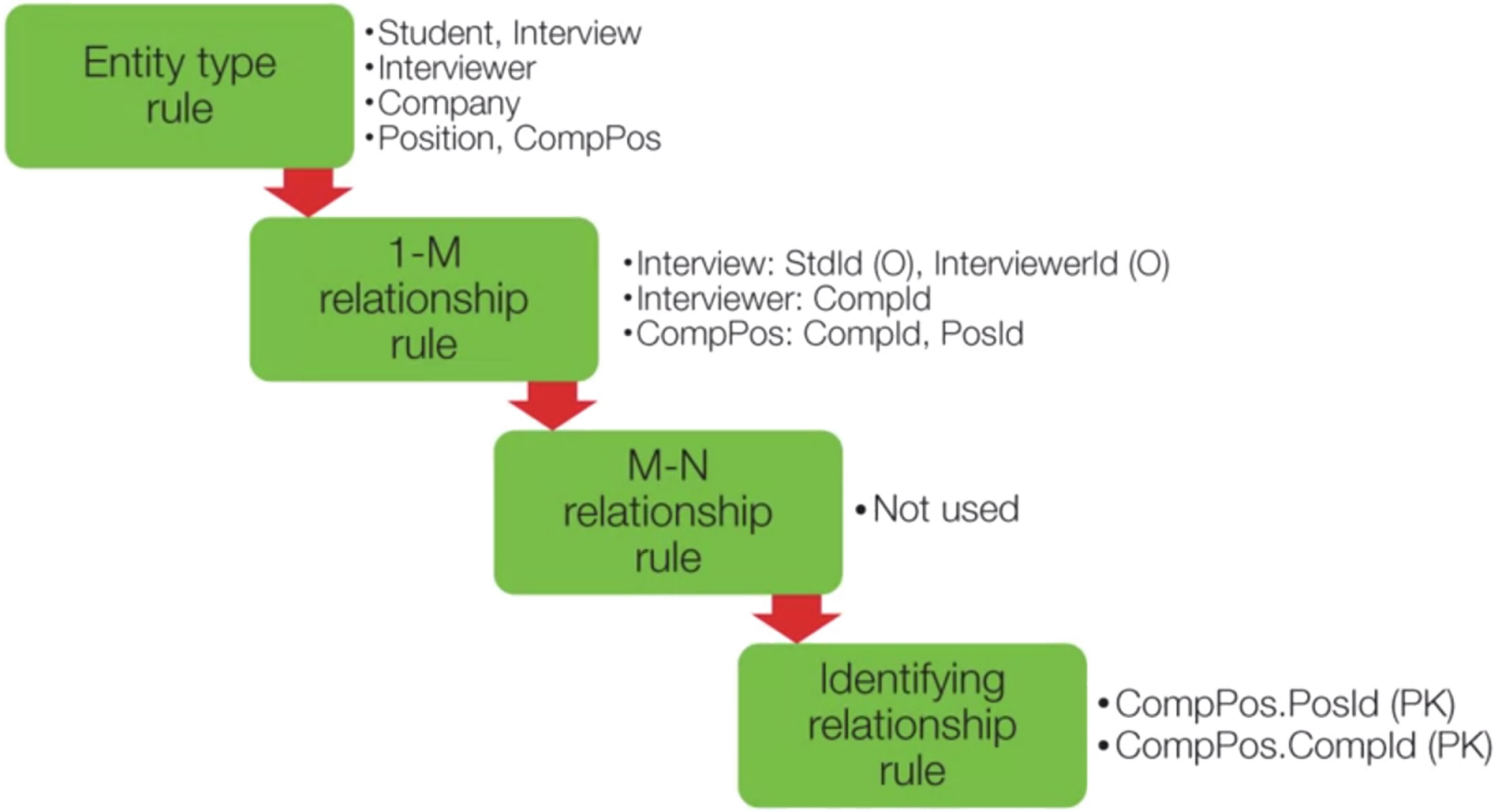
FD Falsification Solution

Big University Table

StdNo	StdCity	StdClass	OfferNo	OffTerm	OffYear	EnrGrade	CourseNo	CrsDesc
S1	SEATTLE	JUN	O1	FALL	2017	3.5	C1	DB
S1	SEATTLE	JUN	O2	FALL	2017	3.3	C2	VB
S2	BOTHELL	JUN	O3	SPRING	2018	3.1	C3	OO
S2	BOTHELL	JUN	O2	FALL	2017	3.4	C2	VB

FD	Falsifications
StdCity → *OfferNo*	(1,2), (3,4)
StdCity → *OffTerm*	(3,4)
StdCity → *EnrGrade*	?,?
StdCity → *CourseNo*	?,?
StdCity → *CrsDesc*	?,?
StdCity → *OffYear*	?,?
StdCity → *StdNo*	None
StdCity → *StdClass*	?,?

Conversion Rule Application



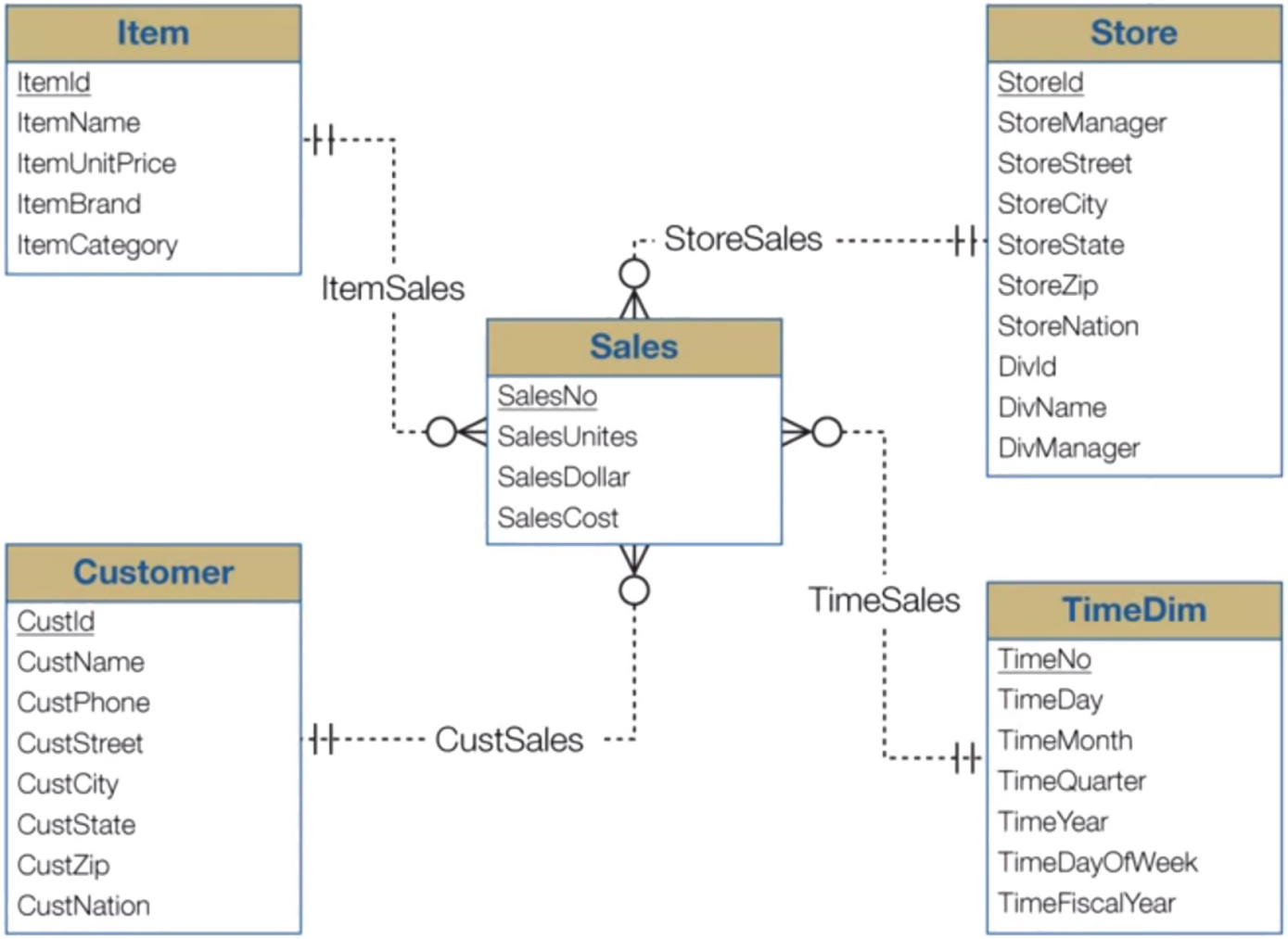
Additional Normalisation

- Only list FDs not implied by PKs

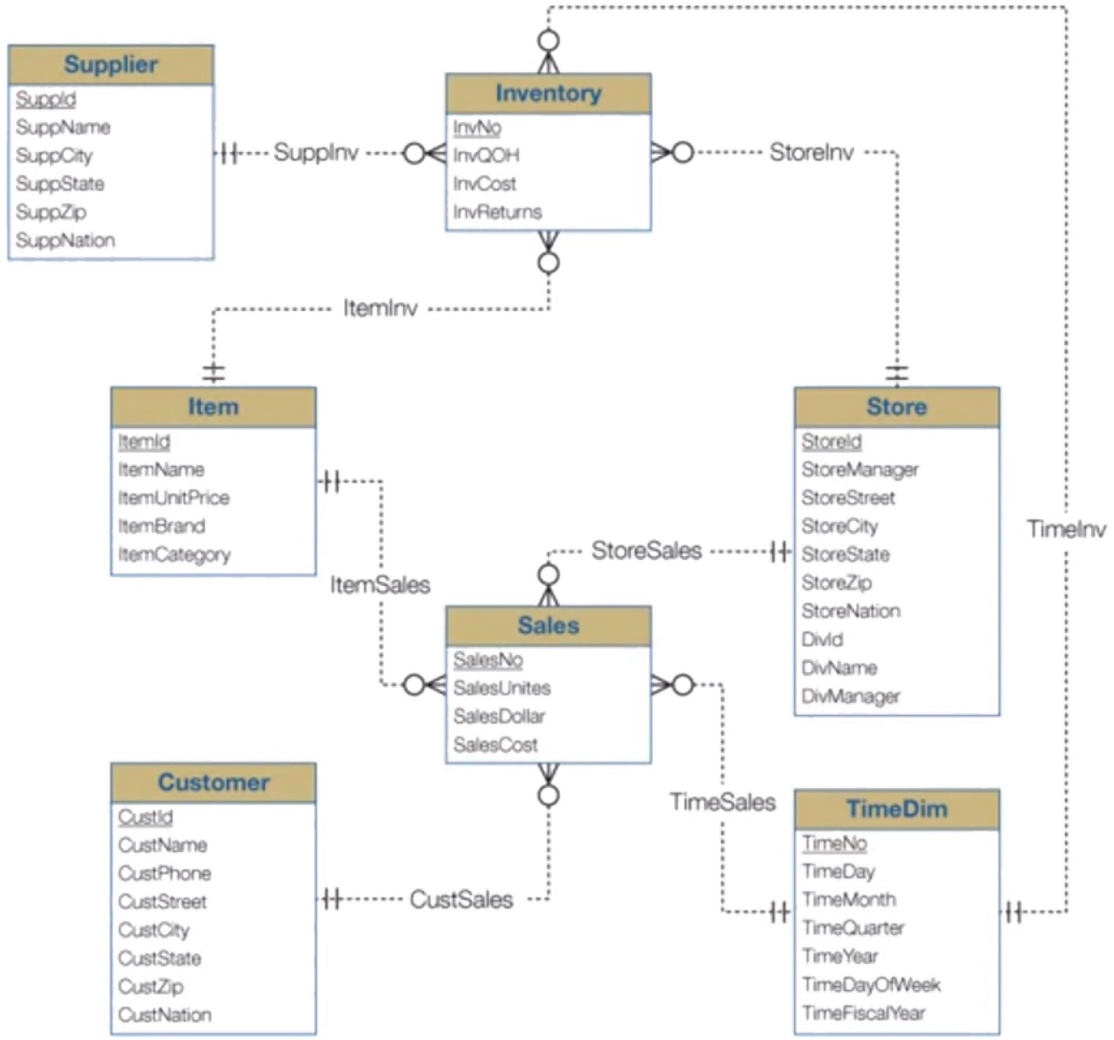
- Additional FDs
  - AdviserNo → AdviserName
  - Possible FD: BldgName, RoomNo → RoomType
  - Possible FD: RoomNo → BldgName, RoomType

What is the origin of the inventory data warehouse design?

Star Schema Example



Constellation Schema Example

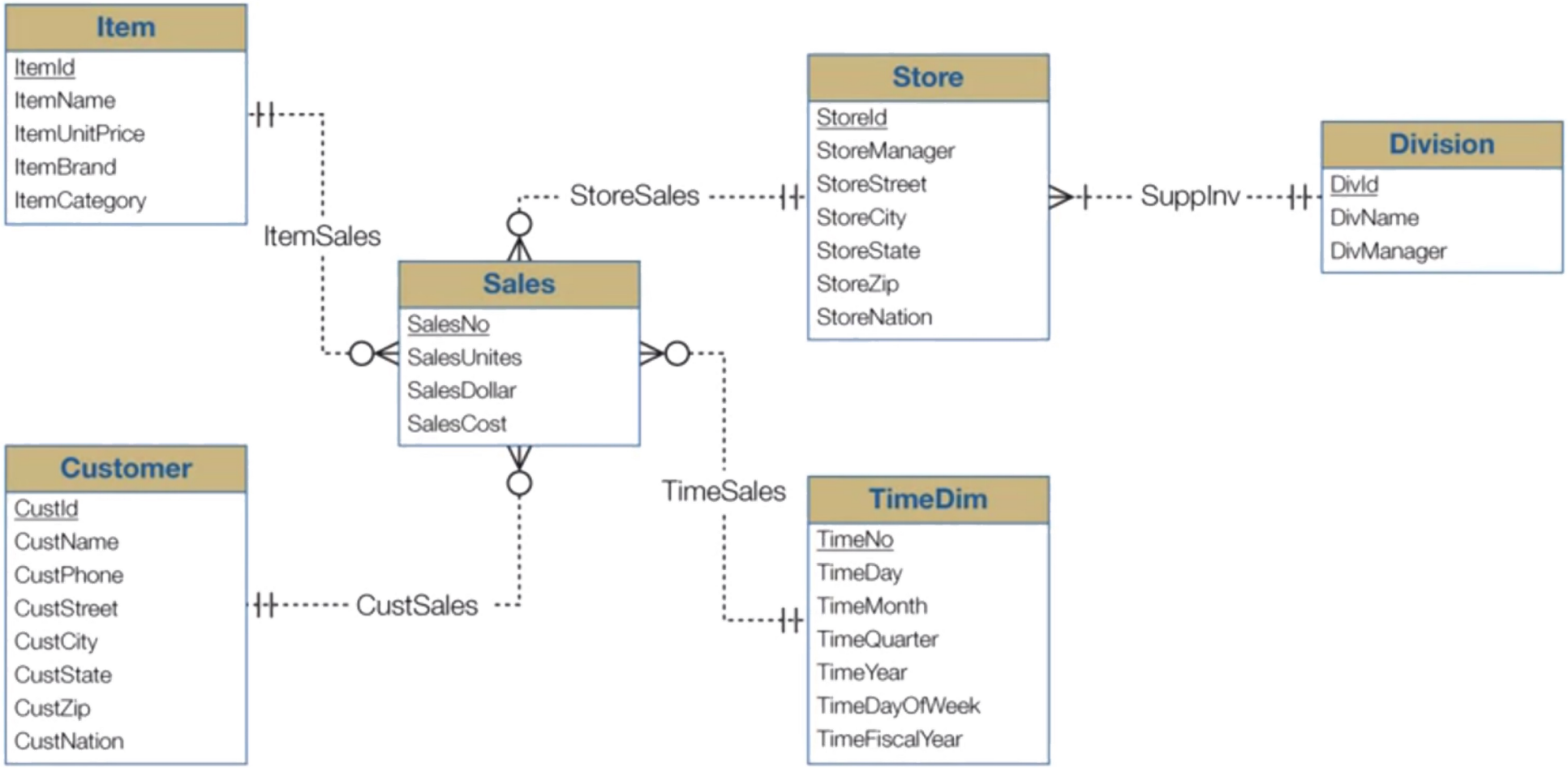


Extended Constellation Schema

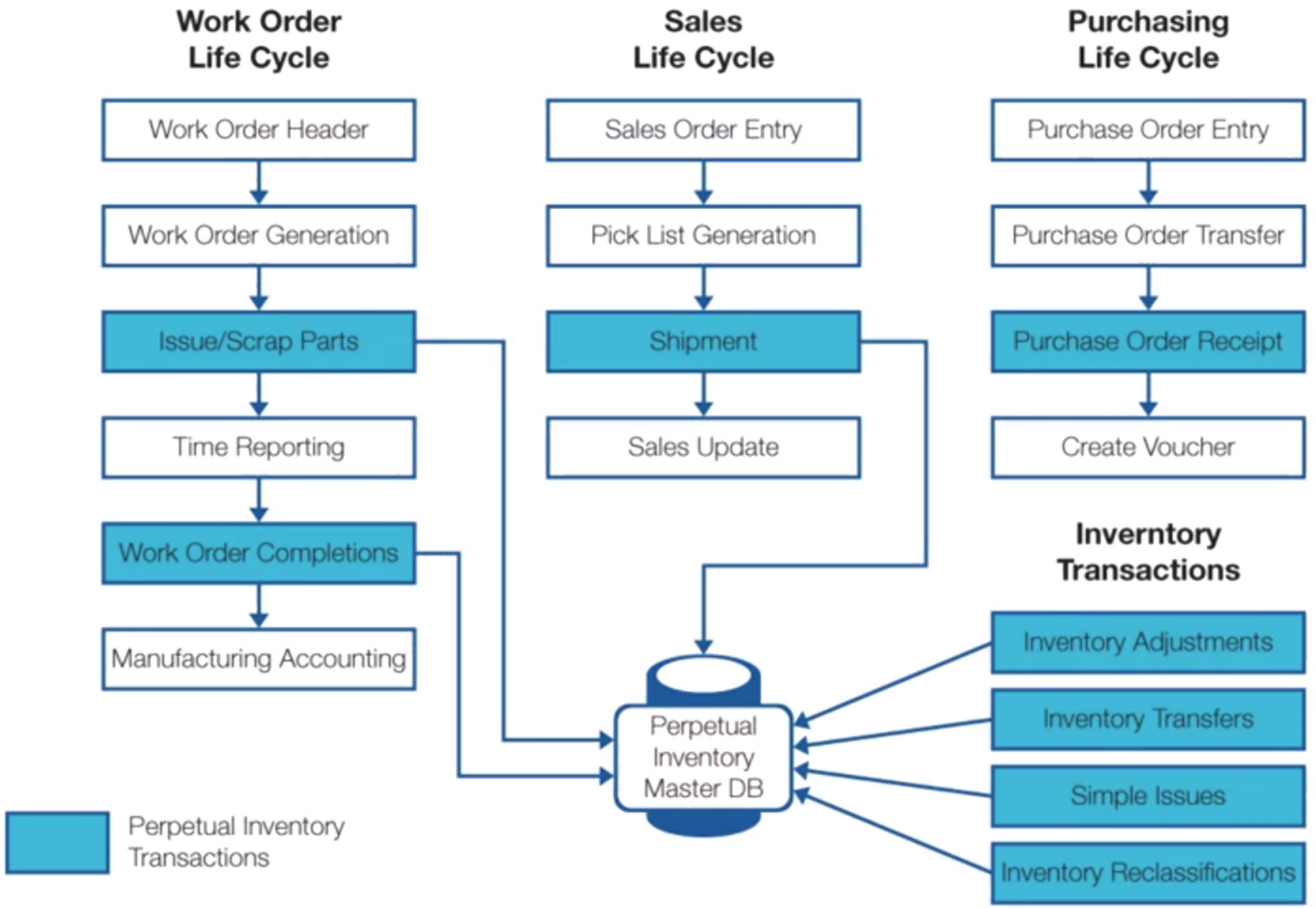
Matrix for a Constellation Schema

Dimensions	Facts		
	Product Sales	Subscription Sales	Inventory
Store	X		
Item	X	X	X
Customer	X	X	
Lead		X	
Package		X	
Format		X	
Supplier			X
TimeDim	X	X	X

Snowflake Schema Example



Inventory Lifecycles



Relational Diagram for the Inventory Data Warehouse