MIS781 Business Intelligence and Database

Module 4: Normalization

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
DEAKIN BUSINESS SCHOOL

NORMALIZATION

- **Normalization**

  - a technique for producing a set of tables with **minimal redundancy** that support the data requirements of an organisation

  - process used to improve the design of relational databases

- **Normal form** - term representing a set of particular conditions (whose purpose is reducing data redundancy) that a table has to satisfy

  - From a lower to a higher normal form, these conditions are increasingly stricter and leave less possibility for redundant data

Benefits of Normalization

Source: Jukic et al. (2020) Database Systems - Introduction to Databases and Data Warehouses, Prospect Press.

Data redundancy and update anomalies

- Major aim of relational database design is to group columns into tables to minimize data redundancy and reduce file storage space required by implemented base tables.
- Problems associated with data redundancy are illustrated by comparing the Staff and Branch tables with the StaffBranch table. See next slide example:

Staff and DistributionCenter tables with StaffDistributionCenter table

Staff

staffNo	name	position	salary	dCenterNo
S1500	Tom Daniels	Manager	48000	D001
S0003	Sally Adams	Assistant	30000	D001
S0010	Mary Martinez	Manager	51000	D002
S3250	Robert Chin	Assistant	33000	D002
S2250	Sally Stern	Manager	48000	D004
S0415	Art Peters	Manager	42000	D003

DistributionCenter

dCenterNo	dAddress	dTelNo
D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
D002	City Center Plaza, Seattle, WA 98122	206-555-6756
D003	14 – 8th Avenue, New York, NY 10012	212-371-3000
D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131

StaffDistributionCenter

staffNo	name	position	salary	dCenterNo	dAddress	dTelNo
S1500	Tom Daniels	Manager	48000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0003	Sally Adams	Assistant	30000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0010	Mary Martinez	Manager	51000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S3250	Robert Chin	Assistant	33000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S2250	Sally Stern	Manager	48000	D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131
S0415	Art Peters	Manager	42000	D003	14 – 8th Avenue, New York, NY 10012	212-371-3000

Data redundancy and update anomalies

- StaffDistributionCenter table has redundant data; the details of a distribution center are repeated for every member of staff.
- In contrast, the details of each distribution center appears only once for each centre in the DistributionCenter table and only the distribution center number (dCenterNo) is repeated in the Staff table, to represent where each member of staff is located.

StaffDistributionCenter

staffNo	name	position	salary	dCenterNo	dAddress	dTelNo
S1500	Tom Daniels	Manager	48000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0003	Sally Adams	Assistant	30000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0010	Mary Martinez	Manager	51000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S3250	Robert Chin	Assistant	33000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S2250	Sally Stern	Manager	48000	D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131
S0415	Art Peters	Manager	42000	D003	14 – 8th Avenue, New York, NY 10012	212-371-3000

Data redundancy and update anomalies

- Tables that contain redundant information may potentially suffer from update anomalies.
- Types of update anomalies include:
  - insertion,
  - deletion,
  - modification.

Solution: NORMALIZATION

- **The normalization process involves examining each table and verifying if it satisfies a particular normal form**
- If a table satisfies a particular normal form, then the next step is to verify if that relation satisfies the next higher normal form
- If a table does not satisfy a particular normal form, actions are taken to convert the table into a set of tables that satisfy the particular normal form

UNF ⟶ 1NF → 2NF → 3NF

- Normalizing to *first normal form* is done on non-relational tables in order to convert them to relational tables
- Normalizing to *subsequent normal forms* (e.g., second normal form, third normal form) improves the design of relational tables that contain redundant information and alleviates the problem of update anomalies

NORMALIZATION

- There are several normal forms, most fundamental of which are:
  - First normal form (1NF)
  - Second normal form (2NF)
  - Third normal form (3NF)

NORMALIZATION (1NF)

- First Normal Form (1NF) - *A table is in 1NF if each row is unique and no column in any row contains multiple values*
  - 1NF states that each value in each column of a table must be a single value from the domain of the column
  - Every relational table is, by definition, in 1NF
  - **Related multivalued columns** - columns in a table that refer to the same real-world concept (entity) and can have multiple values per record
  - Normalizing to 1NF involves eliminating groups of related multi-valued columns

Example: Normalizing a table to 1NF

Non-relational table (not in 1NF).

VET CLINIC CLIENT

ClientID	ClientName	PetNo	PetName	PetType
111	Lisa	1	Tofu	Dog
222	Lydia	1	Fluffy	Dog
		2	JoJo	Bird
		3	Ziggy	Snake
333	Jane	1	Fluffy	Cat
		2	Cleo	Cat

Normalizing the table to 1NF by increasing the number of records

VET CLINIC CLIENT

ClientID	ClientName	PetNo	PetName	PetType
111	Lisa	1	Tofu	Dog
222	Lydia	1	Fluffy	Dog
222	Lydia	2	JoJo	Bird
222	Lydia	3	Ziggy	Snake
333	Jane	1	Fluffy	Cat
333	Jane	2	Cleo	Cat

Example: Normalizing a table to 1NF

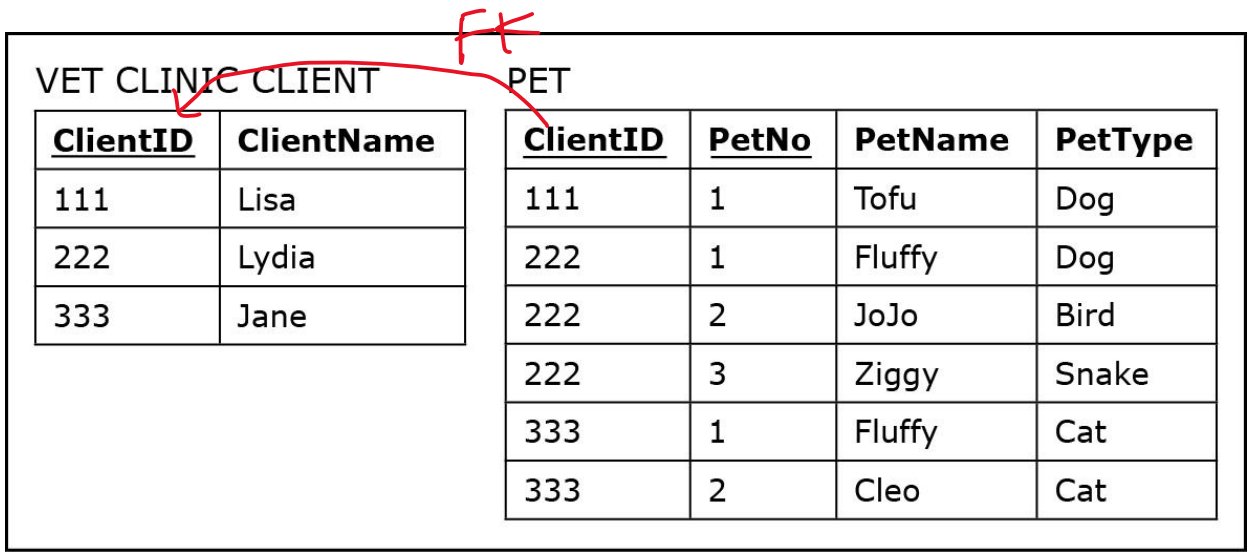
Non-relational table (not in 1NF).

VET CLINIC CLIENT

ClientID	ClientName	PetNo	PetName	PetType
111	Lisa	1	Tofu	Dog
222	Lydia	1	Fluffy	Dog
		2	JoJo	Bird
		3	Ziggy	Snake
333	Jane	1	Fluffy	Cat
		2	Cleo	Cat

Alternatively,

Normalizing the table to 1NF by creating a new, separate table



ClientID	ClientName
111	Lisa
222	Lydia
333	Jane

ClientID	PetNo	PetName	PetType
111	1	Tofu	Dog
222	1	Fluffy	Dog
222	2	JoJo	Bird
222	3	Ziggy	Snake
333	1	Fluffy	Cat
333	2	Cleo	Cat

Example: Normalizing a table to 1NF

Non-relational table (not in 1NF) with two groups of related multivalued columns

VET CLINIC CLIENT

ClientID	ClientName	PetNo	PetName	PetType	HHMember	Name	Relation
111	Lisa	1	Tofu	Dog	1	Joe	Husband
					2	Sally	Daughter
					3	Clyde	Son
222	Lydia	1	Fluffy	Dog	1	Bill	Husband
		2	JoJo	Bird	2	Lilly	Daughter
		3	Ziggy	Snake			
333	Jane	1	Fluffy	Cat	1	Jill	Sister
		2	Cleo	Cat			

Normalizing the table to 1NF

VET CLINIC CLIENT

ClientID	ClientName
111	Lisa
222	Lydia
333	Jane

PET

ClientID	PetNo	PetName	PetType
111	1	Tofu	Dog
222	1	Fluffy	Dog
222	2	JoJo	Bird
222	3	Ziggy	Snake
333	1	Fluffy	Cat
333	2	Cleo	Cat

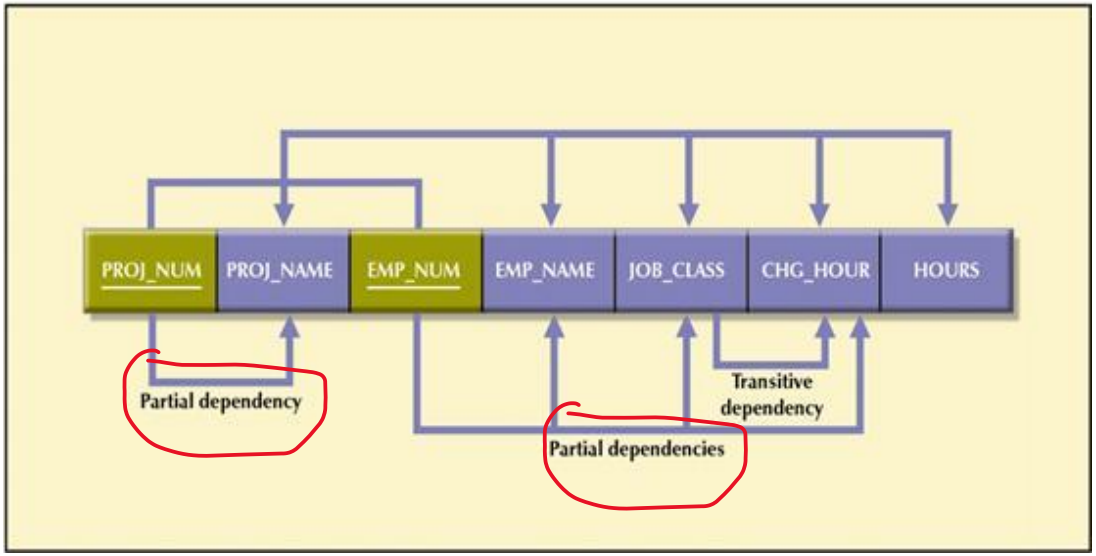
HOUSEHOLD MEMBER

ClientID	HHMember	Name	Relation
111	1	Joe	Husband
111	2	Sally	Daughter
111	3	Clyde	Son
222	1	Bill	Husband
222	2	Lilly	Daughter
333	1	Jill	Sister

NORMALIZATION (2NF)

- **Second Normal Form (2NF)** - *A table is in 2NF if it is in 1NF and if it* ***does not contain partial functional dependencies***
  - If a relation has a single-column primary key, then there is no possibility of partial functional dependencies
    - Such a relation is automatically in 2NF and it does not have to be normalized to 2NF

FIGURE 5.3 A DEPENDENCY DIAGRAM: FIRST NORMAL FORM (1NF)



NORMALIZATION (2NF)

- Second Normal Form (2NF)
  - Normalization of a relation to 2NF creates additional relations for each set of partial dependencies in a relation
    - The primary key of the additional relation is the portion of the primary key that functionally determines the columns in the original relation
    - The columns that were partially determined in the original relation are part of the additional table
  - The original table remains after the process of normalizing to 2NF, but it no longer contains the partially dependent columns

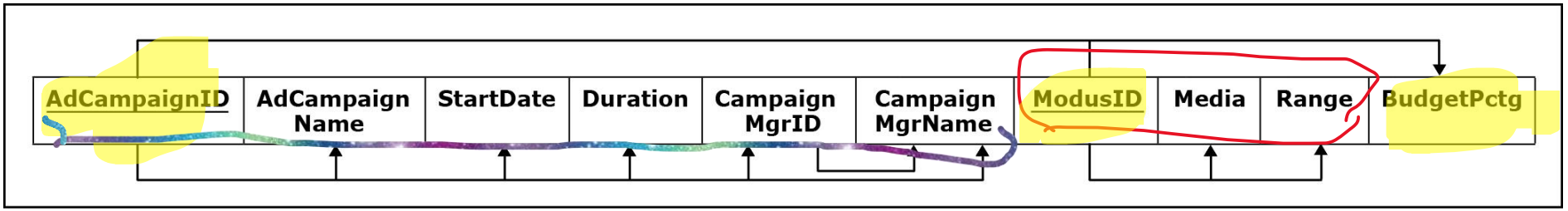
Example Relation AD CAMPAIGN MIX

AD CAMPAIGN MIX

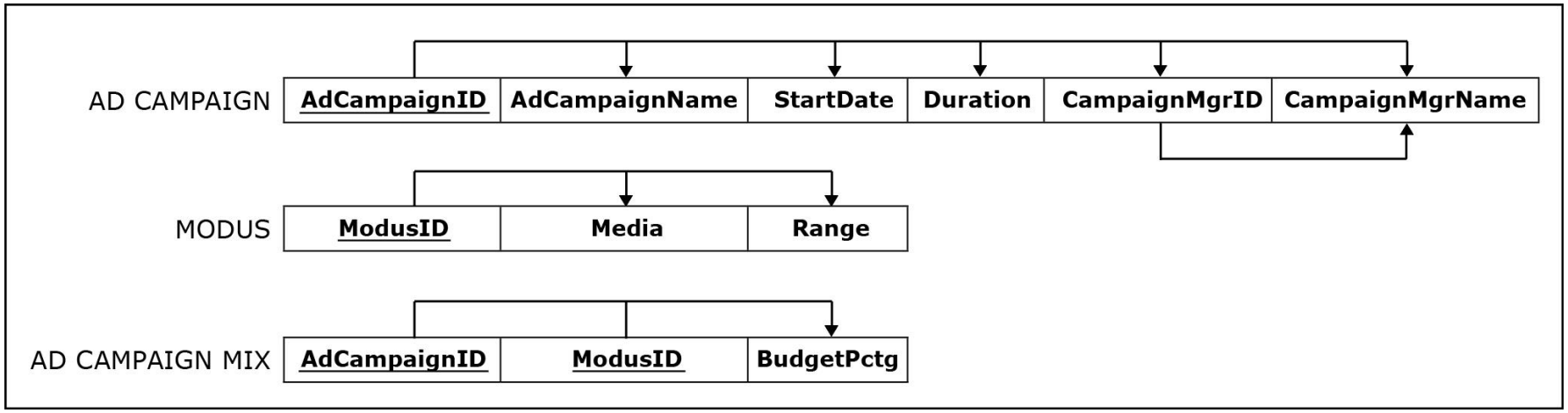
AdCampaignID	**AdCampaign Name**	**StartDate**	**Duration**	**Campaign MgrID**	**Campaign MgrName**	**ModusID**	**Media**	**Range**	**Budget Pctg**
111	SummerFun20	6.6.2020	12 days	CM100	Roberta	1	TV	Local	50%
111	SummerFun20	6.6.2020	12 days	CM100	Roberta	2	TV	National	50%
222	SummerZing20	6.8.2020	30 days	CM101	Sue	1	TV	Local	60%
222	SummerZing20	6.8.2020	30 days	CM101	Sue	3	Radio	Local	30%
222	SummerZing20	6.8.2020	30 days	CM101	Sue	5	Print	Local	10%
333	FallBall20	6.9.2020	12 days	CM102	John	3	Radio	Local	80%
333	FallBall20	6.9.2020	12 days	CM102	John	4	Radio	National	20%
444	AutmnStyle20	6.9.2020	5 days	CM103	Nancy	6	Print	National	100%
555	AutmnColors20	6.9.2020	3 days	CM100	Roberta	3	Radio	Local	100%

Example: Normalizing a table to 2NF

Pressly Ad Agency - relation AD CAMPAIGN MIX

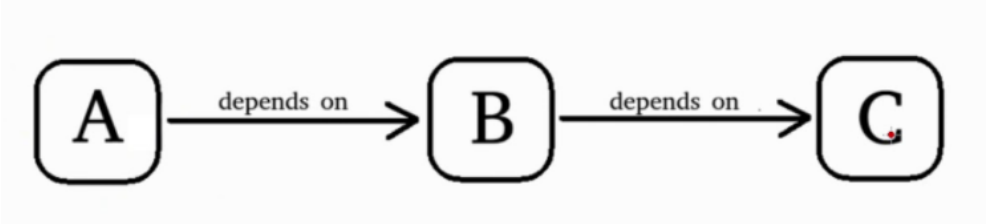


Pressly Ad Agency example - normalized to 2NF



NORMALIZATION (3NF)

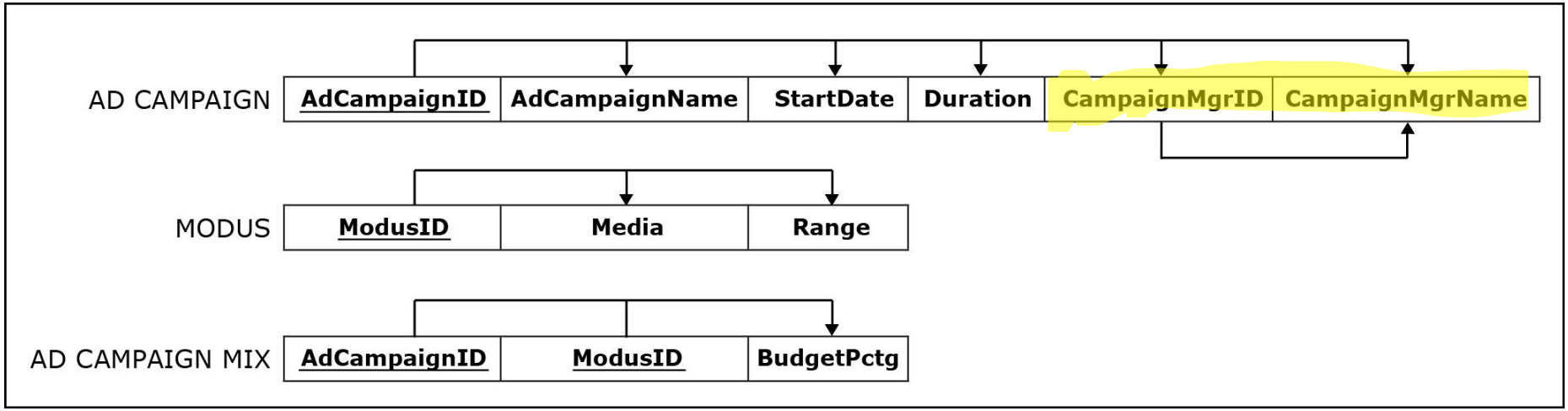
- **Third Normal Form (3NF)** - *A table is in 3NF if it is in 2NF and if it **does not contain transitive functional dependencies.***
- Eg: A is transitive dependent on C:



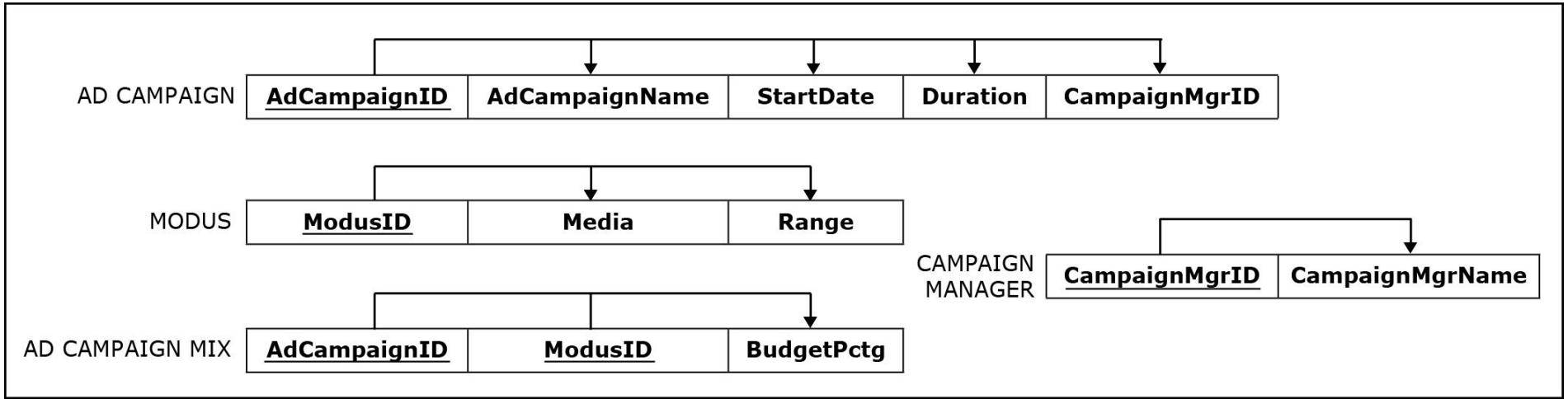
  - Normalization of a relation to 3NF creates additional relations for each set of transitive dependencies in a relation.
    - The primary key of the additional relation is the nonkey column (or columns) that functionally determined the nonkey columns in the original relation
    - The nonkey columns that were transitively determined in the original relation are part of the additional table.
  - The original table remains after normalizing to 3NF, but it no longer contains the transitively dependent columns

Example: Normalizing a table to 3NF

Pressly Ad Agency example - normalized to 2NF



Pressly Ad Agency example - normalized to 3NF



Pressly Ad Agency relation AD CAMPAIGN MIX – not normalized, prone to update anomalies

AD CAMPAIGN MIX

AdCampaignID	**AdCampaign Name**	**StartDate**	**Duration**	**Campaign MgrID**	**Campaign MgrName**	**ModusID**	**Media**	**Range**	**Budget Pctg**
111	SummerFun20	6.6.2020	12 days	CM100	Roberta	1	TV	Local	50%
111	SummerFun20	6.6.2020	12 days	CM100	Roberta	2	TV	National	50%
222	SummerZing20	6.8.2020	30 days	CM101	Sue	1	TV	Local	60%
222	SummerZing20	6.8.2020	30 days	CM101	Sue	3	Radio	Local	30%
222	SummerZing20	6.8.2020	30 days	CM101	Sue	5	Print	Local	10%
333	FallBall20	6.9.2020	12 days	CM102	John	3	Radio	Local	80%
333	FallBall20	6.9.2020	12 days	CM102	John	4	Radio	National	20%
444	AutmnStyle20	6.9.2020	5 days	CM103	Nancy	6	Print	National	100%
555	AutmnColors20	6.9.2020	3 days	CM100	Roberta	3	Radio	Local	100%

Pressly Ad Agency example—normalized relations with data

AD CAMPAIGN

AdCampaignID	AdCampaignName	StartDate	Duration	CampaignMgrID
111	SummerFun20	6.6.2020	12 days	CM100
222	SummerZing20	6.8.2020	30 days	CM101
333	FallBall20	6.9.2020	12 days	CM102
444	AutmnStyle20	6.9.2020	5 days	CM103
555	AutmnColors20	6.9.2020	3 days	CM100

CAMPAIGN MANAGER

CampaignMgrID	CampaignMgrName
CM100	Roberta
CM101	Sue
CM102	John
CM103	Nancy

MODUS

ModusID	Media	Range
1	TV	Local
2	TV	National
3	Radio	Local
4	Radio	National
5	Print	Local
6	Print	National

AD CAMPAIGN MIX

AdCampaignID	ModusID	BudgetPctg
111	1	50%
111	2	50%
222	1	60%
222	3	30%
222	5	10%
333	3	80%
333	4	20%
444	6	100%
555	3	100%

Another normalization example: Normalizing a table to **2NF**

Central Plane University - relation RECRUITING

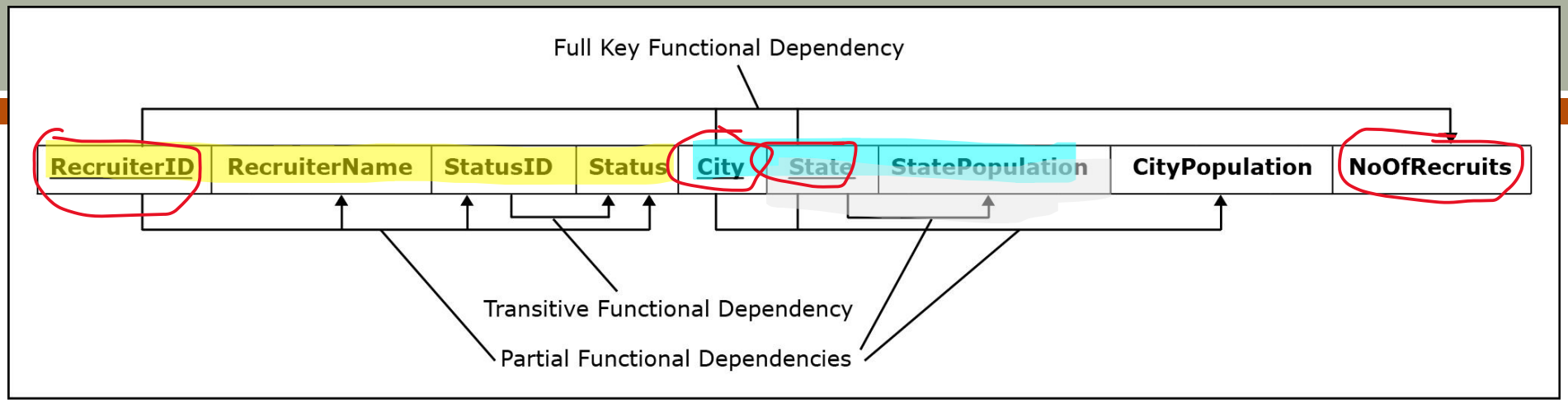
Central Plane University relation RECRUITING – not normalized, prone to update anomalies

RECRUITING

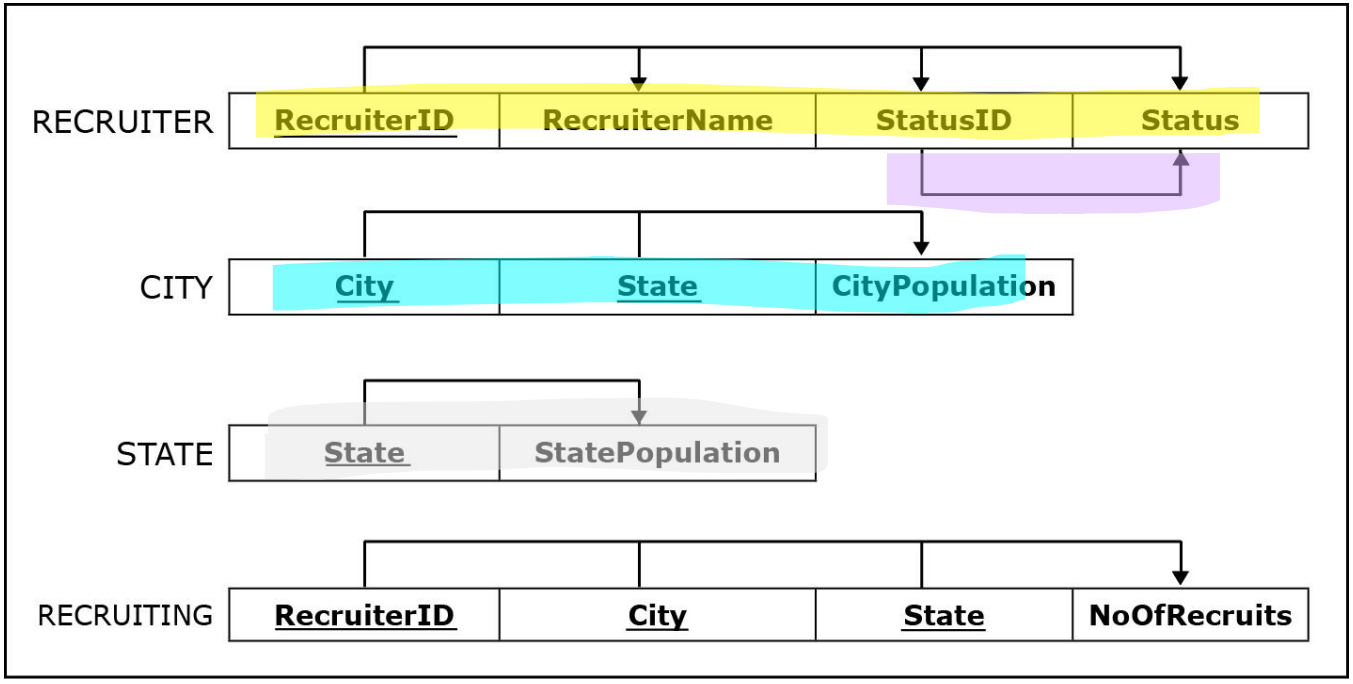
RecruiterID	**RecruiterName**	**StatusID**	**Status**	**City**	**State**	**State Population**	**City Population**	**NoOfRecruits**
R1	Katy	IF	Internal Full Time	Portland	ME	1,350,000	70,000	11
R1	Katy	IF	Internal Full Time	Grand Rapids	MI	9,900,000	190,000	20
R2	Abby	IP	Internal Part Time	Rockford	IL	12,900,000	340,000	17
R3	Jana	CN	Contractor	Spokane	WA	6,800,000	210,000	8
R3	Jana	CN	Contractor	Portland	OR	3,900,000	600,000	30
R3	Jana	CN	Contractor	Eugene	OR	3,900,000	360,000	20
R4	Maria	IF	Internal Full Time	Rockford	IL	12,900,000	340,000	14
R4	Maria	IF	Internal Full Time	Grand Rapids	MN	5,400,000	11,000	9
R5	Dan	CN	Contractor	Grand Rapids	MI	9,900,000	190,000	33

Another normalization example: Normalizing a table to 2NF

Central Plane University - relation RECRUITING

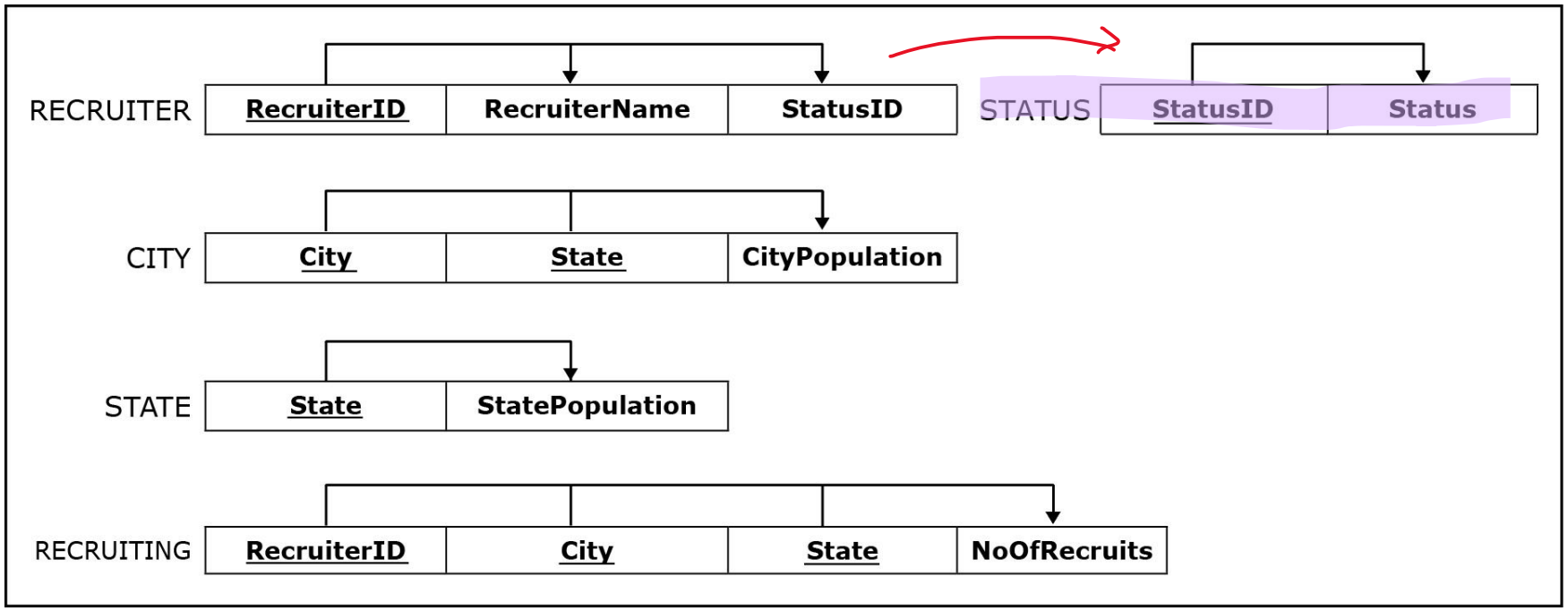


Central Plane University example - normalized to 2NF



Another normalization example: Normalizing a table to 3NF

Central Plane University example - normalized to 3NF (5 tables)



Central Plane University relation RECRUITING – not normalized, prone to update anomalies

RECRUITING

RecruiterID	RecruiterName	StatusID	Status	City	State	State Population	City Population	NoOfRecruits
R1	Katy	IF	Internal Full Time	Portland	ME	1,350,000	70,000	11
R1	Katy	IF	Internal Full Time	Grand Rapids	MI	9,900,000	190,000	20
R2	Abby	IP	Internal Part Time	Rockford	IL	12,900,000	340,000	17
R3	Jana	CN	Contractor	Spokane	WA	6,800,000	210,000	8
R3	Jana	CN	Contractor	Portland	OR	3,900,000	600,000	30
R3	Jana	CN	Contractor	Eugene	OR	3,900,000	360,000	20
R4	Maria	IF	Internal Full Time	Rockford	IL	12,900,000	340,000	14
R4	Maria	IF	Internal Full Time	Grand Rapids	MN	5,400,000	11,000	9
R5	Dan	CN	Contractor	Grand Rapids	MI	9,900,000	190,000	33

Central Plane University example - normalized relations with data (redundancy eliminated and update anomalies resolved)

RECRUITER

RecruiterID	RecruiterName	StatusID
R1	Katy	IF
R2	Abra	IP
R3	Jana	CN
R4	Maria	IF
R5	Dan	CN

STATE

State	StatePopulation
ME	1,350,000
MI	9,900,000
IL	12,900,000
WA	6,800,000
OR	3,900,000
MN	5,400,000

STATUS

StatusID	Status
CN	Contractor
IF	Internal Full Time
IP	Internal Part Time

RECRUITING

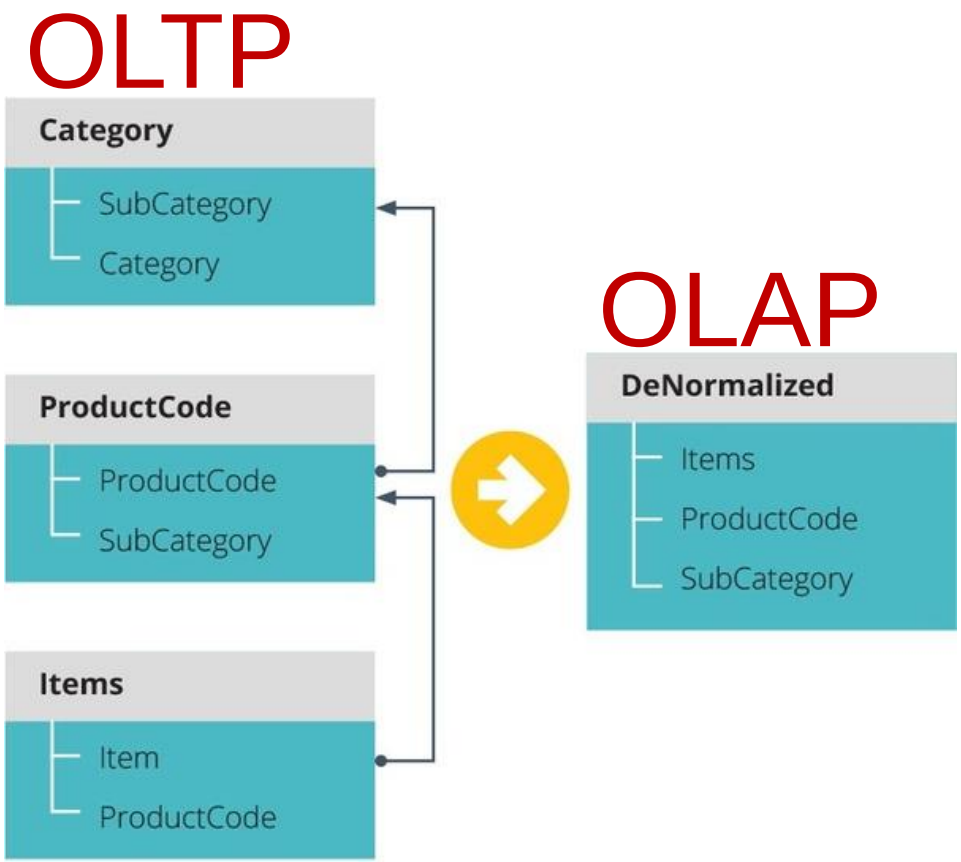
RecruiterID	City	State	NoOfRecruits
R1	Portland	ME	11
R1	Grand Rapids	MI	20
R2	Rockford	IL	17
R3	Spokane	WA	8
R3	Portland	OR	30
R3	Eugene	OR	20
R4	Rockford	IL	14
R4	Grand Rapids	MN	9
R5	Grand Rapids	MI	33

CITY

City	State	CityPopulation
Portland	ME	70,000
Grand Rapids	MI	190,000
Rockford	IL	340,000
Spokane	WA	210,000
Portland	OR	600,000
Eugene	OR	360,000
Grand Rapids	MN	11,000

Normalization vs Denormalization

- **Normalization** is the technique of dividing the data into multiple tables to reduce data redundancy and to achieve data integrity.
- **Denormalization** is the technique of combining the data into a single table to make data retrieval faster.



DENORMALIZATION

- **Denormalization** - reversing the effect of normalization by joining normalized relations into a relation that is not normalized, in order to improve query performance
  - The data that resided in fewer relations prior to normalization is spread out across more relations after normalization
  - This has an effect on the performance of data retrievals
  - Denormalization can be used in dealing with the normalization vs. performance issue
- Denormalization is not a default process that is to be undertaken in all circumstances
  - Instead, denormalization should be used judiciously, after analysing its costs and benefits

Denormalization example: Quicker retrieval

Pressly Ad Agency example—a retrieval of data

RETRIEVED DATA

AdCampaignID	*AdCampaign Name*	*Campaign MgrID*	*Campaign MgrName*	*ModusID*	*Media*	*Range*	*BudgetPctg*
111	SummerFun20	CM100	Roberta	1	TV	Local	50%
111	SummerFun20	CM100	Roberta	2	TV	National	50%
222	SummerZing20	CM101	Sue	1	TV	Local	60%
222	SummerZing20	CM101	Sue	3	Radio	Local	30%
222	SummerZing20	CM101	Sue	5	Print	Local	10%
333	FallBall20	CM102	John	3	Radio	Local	80%
333	FallBall20	CM102	John	4	Radio	National	20%
444	AutmnStyle20	CM103	Nancy	6	Print	National	100%
555	AutmnColors20	CM100	Roberta	3	Radio	Local	100%

OLTP DATA QUALITY

- **OLTP Data quality**
  - The data in a database is considered of high quality if it correctly and non-ambiguously reflects the real-world it is designed to represent
  - Data quality characteristics
    - Accuracy
    - Uniqueness
    - Completeness
    - Consistency
    - Timeliness
    - Conformity

OLTP DATA QUALITY

- **Accuracy** - the extent to which data correctly reflects the real-world instances it is supposed to depict (eg. Misspelling customer name)
- **Completeness** - the degree to which all the required data is present in the data collection
- **Uniqueness** - requires each real-world instance to be represented only once in the data collection
  - The uniqueness data quality problem is sometimes also referred to as data duplication (e.g. 2 records refer to the same customers)

OLTP DATA QUALITY

- **Consistency** - the extent to which the data properly conforms to and matches up with the other data (e.g. profit from sales vs account dept)
- **Timeliness** - the degree to which the data is aligned with the proper time window in its representation of the real world (eg. Delivery of parcel)
  - Typically, timeliness refers to the “freshness” of the data
- **Conformity** - the extent to which the data conforms to its specified format (e.g. $99.9 or $99,9).

Data quality – Example

A message reporting the head count of the managers in the Albritco company

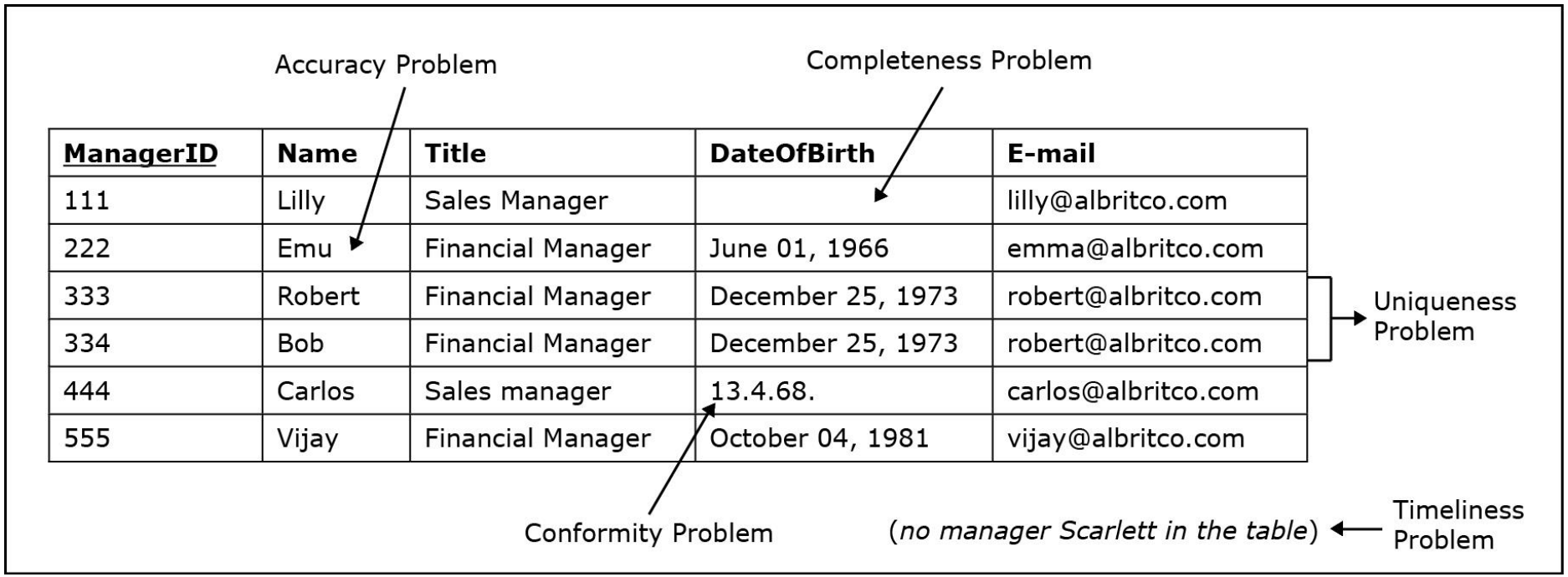
To: Albritco Board of Directors
Subject: Strategic Planning Goal Achieved (Confidential)

We are happy to report that the strategic goal of having an equal number of sales and financial managers has been achieved as shown below:

Manager's Head Count Uniformity: Goal Achieved	
Number of Sales Managers:	3
Number of Financial Managers:	3

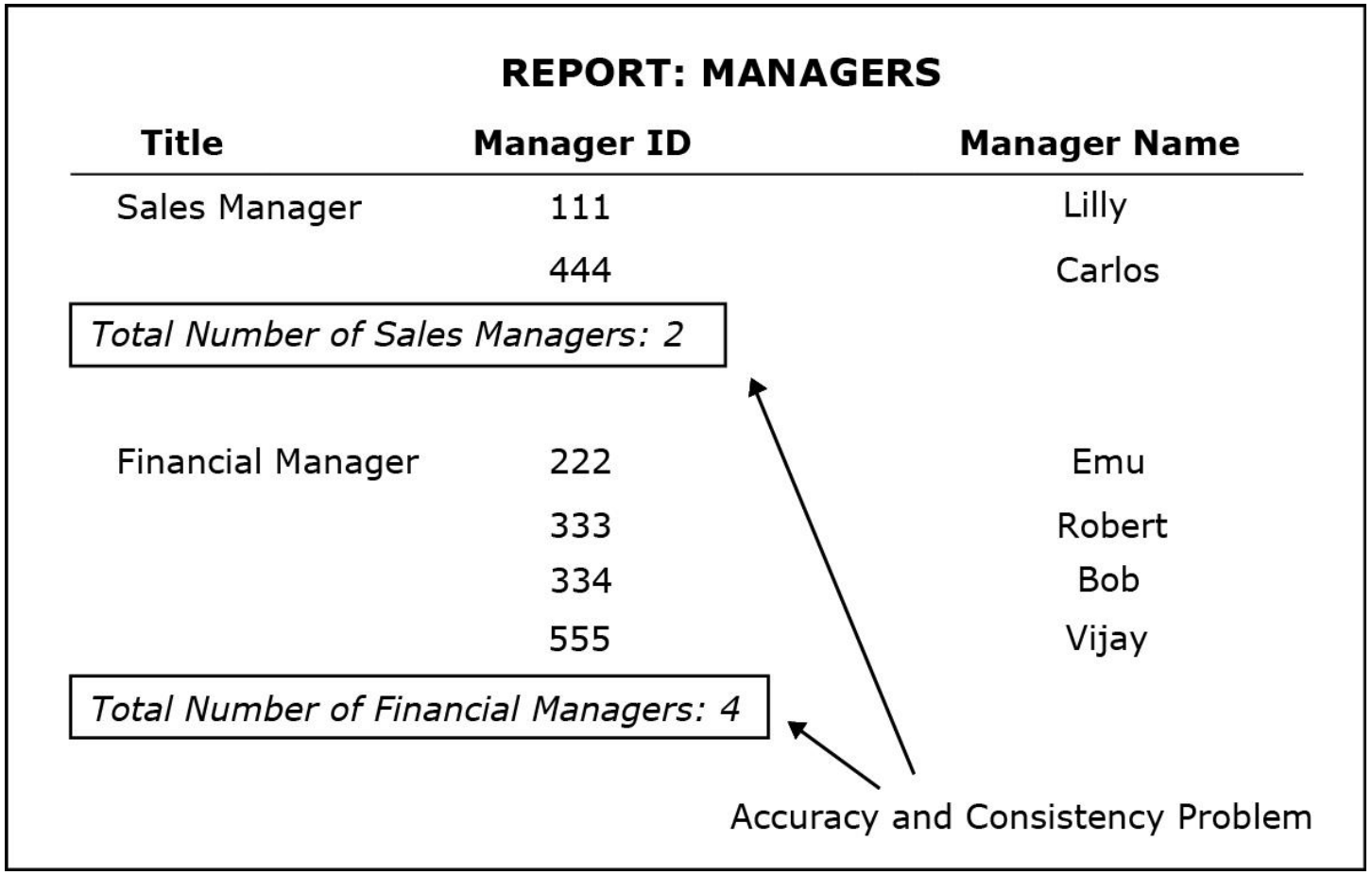
Data quality – Example

Albritco database relation with data quality issues



Data quality – Example

A report based on the Albritco relation with data quality issues



Data quality – Example

Albritco database relation with the data quality issues resolved

ManagerID	**Name**	**Title**	**DateOfBirth**	**E-mail**
111	Lilly	Sales Manager	May 21, 1971	lilly@albritco.com
222	Emma	Financial Manager	June 01, 1966	emma@albritco.com
333	Robert	Financial Manager	December 25, 1973	robert@albritco.com
444	Carlos	Sales Manager	April 13, 1968	carlos@albritco.com
555	Vijay	Financial Manager	October 04, 1981	vijay@albritco.com
666	Scarlett	Sales Manager	July 17, 1984	scarlett@albritco.com

Data quality – Example

A report based on the Albritco relation with the data quality issues resolved

REPORT: MANAGERS

Title	Manager ID	Manager Name
Sales Manager	111	Lilly
	444	Carlos
	666	Scarlett
Total Number of Sales Managers: 3		
Financial Manager	222	Emma
	333	Robert
	555	Vijay
Total Number of Financial Managers: 3		

Assignment Discussion

- Assignment – see sample outline & normalization tutorial