

CIS 9440 - Data Warehousing and Analytics

Class #1




Week 1 Class Overview:

1. What is Data Warehousing?
2. Course Introduction
3. CIS 9340 Review
4. Intro to BigQuery



Week 1 Class Overview:

1. What is Data Warehousing?
 2. Course Introduction
 3. CIS 9340 Review
 4. Intro to BigQuery
- 

The background is a solid teal color. In the top-left corner, there are three vertical bars of varying heights, all in a lighter shade of teal. In the bottom-right corner, there are four vertical bars of increasing height from left to right, also in the same lighter shade of teal.

What is Data Warehousing?



What is Data Warehousing?

Data Warehousing is the process of staging an organization's' data for optimal access, reporting, and analysis.

- With proficient Data Warehousing, any member of an organization can quickly and accurately make an informed, data-driven business decision using a Business Intelligence (BI) tool.
- How is this different than data capture and storage? (CIS 9340)



More Details...

Organizations (should) have 2 sources of data:

1. Operational Database - where day-to-day business transactions are stored/inserted.
 - a. Fast and repetitive
 - b. Normalized
 - c. For data capture, not analytics



More Details...

2. Data Warehouse/Business Intelligence - data for reporting, querying, and analytics.

- Large and historical queries
- Built for making business decisions, analytics
- Not updated at every transaction



More Details...

Data Warehousing is the process preparing data from the Operational Database (source 1) for proper Business Intelligence (source 2)

In simplest terms, the Operational System is where data goes in and the Data Warehouse is where data goes out.

How do you accomplish Data Warehousing?

Data Warehousing is comprised of **3 layers**:

1. Data Integration
2. Data Warehouse
3. Business Intelligence

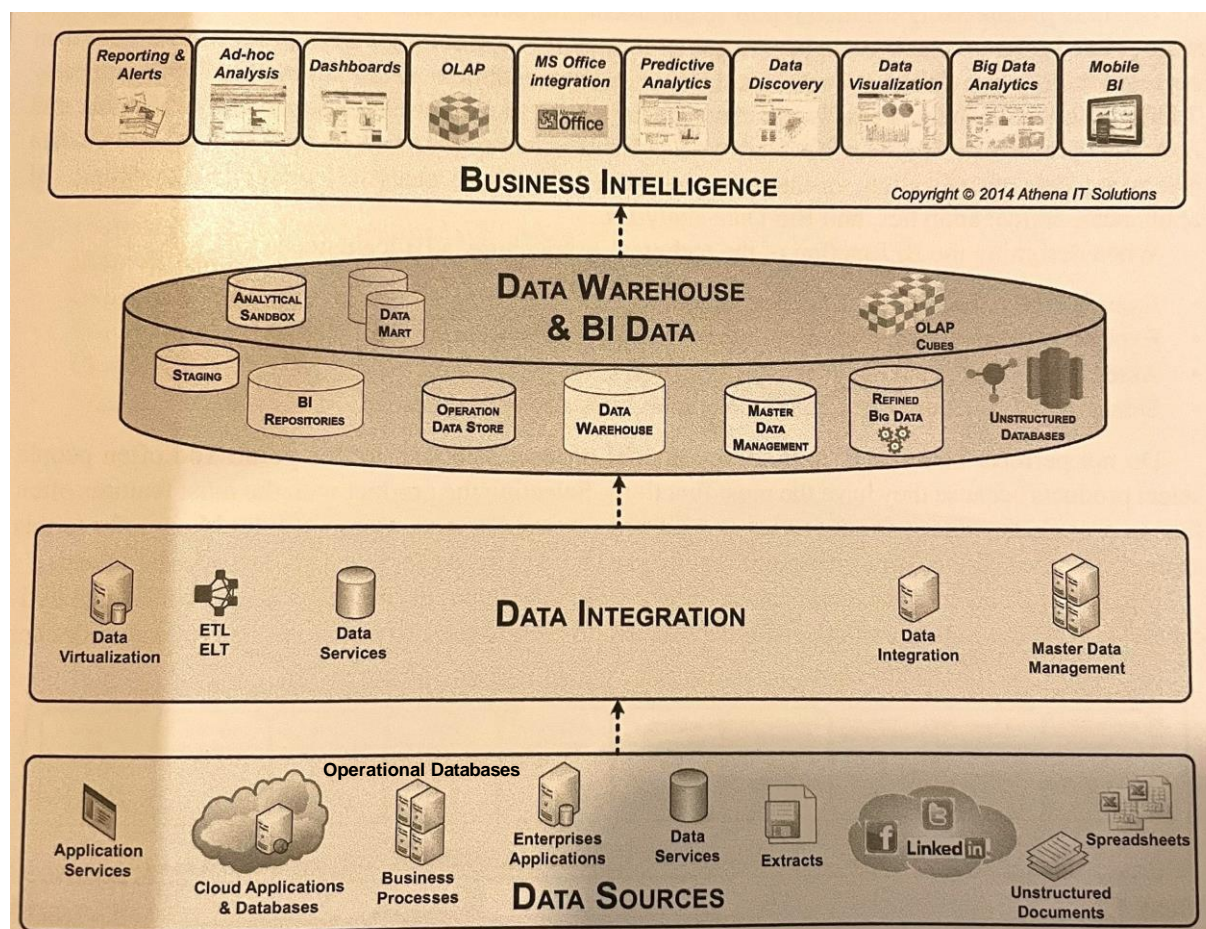


FIGURE 4.6

BI technical architecture categories.



Data Warehousing Example

Suppose you're a Movie Streaming Service, you have:

1. Operational Database

- a. which movies were streamed
- b. reviews were given
- c. when subscriptions started/ended/cancelled
- d. times customer support was contacted
- e. records of customer payments


2. Business Intelligence

- a. How many subscriptions did we gain/lose this month?
- b. How did that compare to this month last year?
- c. Which movies should we remove from our service?
- d. Which movie genres led to most subscriber gains this year?



Who needs Data Warehousing?

- Any organizations that want to utilize **Business Intelligence platform(s)** for analyses, reports, dashboards, scorecards, predictive models, etc
- Any organization with multiple data sources
 - As an example, a Movie Subscription Service has revenue, customer support, payroll, streaming stats, online reviews, social media data/metrics, catalog of movies, etc



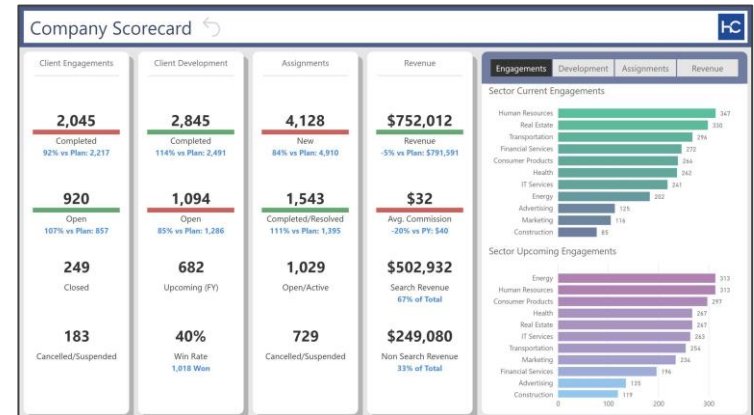
What's the difference between Data Warehousing and a Data Warehouse?

- **Data Warehouse**: physical server that holds the staged data, and connects to the BI Platform
- **Data Warehousing**: the process of staging an organization's' data for optimal access, reporting, and analysis.



How do End Users interact with a Data Warehouse?

- BI Applications for an organization's Analysts
- Company Dashboards/Scorecards
- Recurring Company Reports (to an org/division)
- Structured Tables for Predictive Analytics/Ad-hoc Analyses
- How have you used a Data Warehouse?



Short Quiz

Class Question 1:

Which system do you think holds more historical data? Why?

- A. Operational System (Data Capture)
- B. Data Warehouse (DW/BI)

Class Question 2:

In which system is it more important to have accurate data? Why?

- A. Operational System (Data Capture)
- B. Data Warehouse (DW/BI)



Why Learn Data Warehousing?

- Relevant: Data Warehousing is leveraged (or needed) at nearly every organization with multiple data sources, from startups to Fortune 100 companies
 - Your understanding of this topic is valuable to many organizations
- Necessary: Data Warehousing is both complex and essential
 - Not all Analysts will understand the components and logic of Data Warehousing, although all should
- Empowering: You can (and will) bring an analytical idea to life with a Data Warehousing Project (Final Project of this course)



What will your Final Project look like?

- Capture public data from multiple sources
- Properly store, transform, and stage the data with Data Warehousing, using Dimensional Modeling
- Create a public-facing dashboard/visualization that provides needed insights about your data

I encourage you to start thinking about datasets and questions that interest you! Check out public datasets like Gapminder data, NYC Open Data, Google's public datasets, etc. How could you uniquely combine datasets to create insights?

The background is a solid teal color. In the top-left corner, there are three vertical bars of varying heights, all in a lighter teal shade. In the bottom-right corner, there are four vertical bars of increasing height from left to right, also in the same lighter teal shade.

Week 1 Class Overview:

1. What is Data Warehousing?
2. Course Introduction
3. CIS 9340 Review
4. Intro to BigQuery



Course Introduction



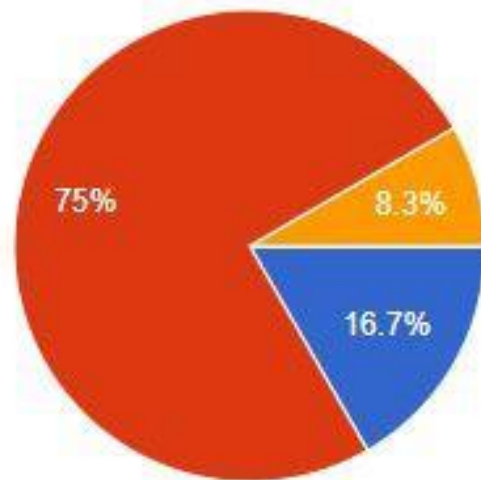
About Me

- Academic background in Mathematics and Data Science
- Data Scientist
 - Expertise: python, SQL, EDA, NLP, regression, Executive Insights, BI Reporting, Customer Success, Retail
 - Worked at large companies and startups (technical interview advice)

About You?

When do you expect to graduate?

12 responses

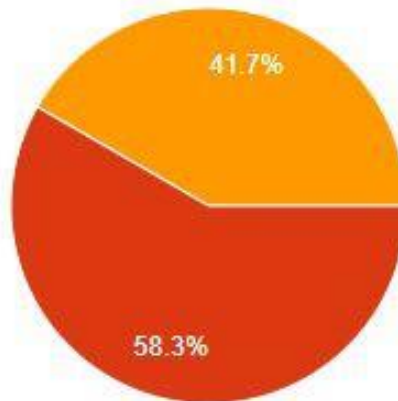


- Fall 2022
- Spring 2023
- Fall 2023
- Other

About You?

Do you have experience with SQL?

12 responses

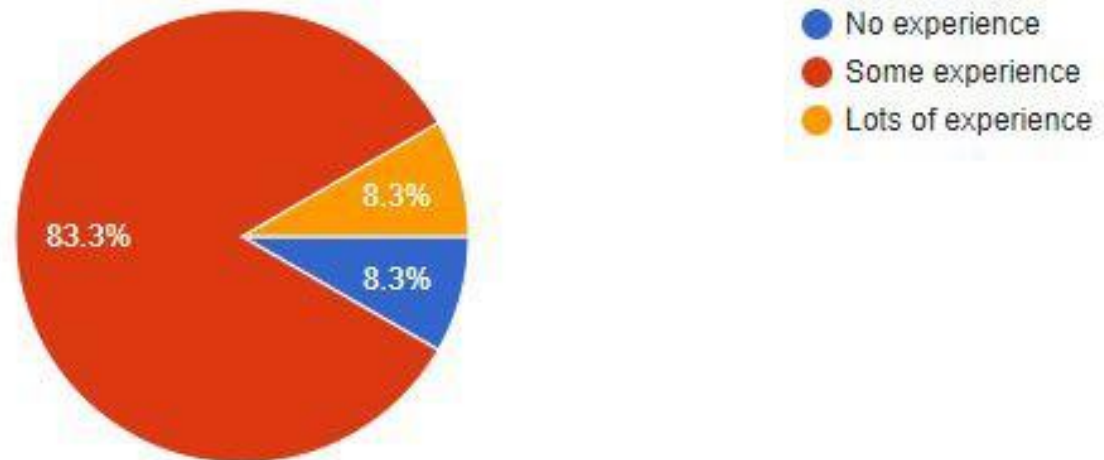


- No experience
- Some experience, only from CIS 9340
- Some experience
- Lots of experience

About You?

Do you have experience programming in python?

12 responses



Student Introductions

- What interested you to take Data Warehousing for Analytics?
- What field/industry do you hope to work in? Ideal job title?
- Tell the class about yourself, any interesting jobs, internships, hobbies?



Course Objective (as an OKR)

from Measure What Matters by John Doerr

- **Objective:** CIS 9440 students will design and implement an *industry-relevant* Data Warehousing solution with a BI Platform (for their *personal portfolio*).
 - Key Result 1: Students will build original conceptual, logical, and physical data models
 - Key Result 2: Students will build an applicable Data Architecture for a Data Warehousing solution
 - Key Result 3: Students will implement an Data Integration process with either ETL
 - Key Result 4: Students will connect a BI Platform to a Data Warehouse



Course Topics

- Data Warehouse Modeling, **Dimensional Modeling**
- Data Warehouse Architecture
- Relational Design and SQL
- **Extract, Transform, and Load (ETL)**, and ELT
- **BI Applications Design and Development**
- NoSQL Databases
- *Data Warehousing in the today's industry*



Expectations during each Class

- Attendance is mandatory, please let me know before class if you cannot attend
- Please bring your computer to all classes.
 - If you do not have a laptop, please work with BCTC to borrow a laptop (more details in syllabus)
- Come to class expecting to participate - nearly all classes will have exercises/workshops
- Ask questions, relating to real-world examples if helpful
- The modality of this class is HyFlex (details on next slide)



What is HyFlex?

- Every class will be synchronously hosted in-person and via Zoom.
- You may choose which classes you attend in-person; you must attend at least 5 classes in-person
- When you attend via Zoom, write your name in the Zoom Chat in the first 5 minutes to receive credit for attendance
- The Midterm Exam and Final Exam do not count as in-person attendance
- Questions about this course modality/HyFlex?



Data Warehousing Tools

- Data Warehousing tools used this semester include, but are not limited to:
 - [Lucidchart](#) - create data models
 - [Google BigQuery](#) - build and manage a Data Warehousing solution
 - [python](#) - for ETL pipelines
 - Google Cloud Functions, Google Cloud Scheduler - schedule custom ETL jobs into BigQuery
 - [Tableau](#) - BI application to create data visualizations and dashboards
- If there is a particular tool you would like to use in this course, let me know



Course Prerequisite

- **CIS 9340**
 - SQL
 - ER Diagrams
 - Normalization
 - Data Modeling - Conceptual, Logical, Physical



Course Structure

- 15 Weeks
- 2 Homework Assignments
- 2 Exams: Midterm Exam and Final Exam
- 1 Final Project
 - Broken into 5 deliverable Milestones that will have due dates throughout the semester
 - Final Project can be completed individually or as a group
- Reading assigned every week



Grading Summary

Assessment	Percentage of Final Grade
Data Warehousing Project	35%
Homeworks	35%
Final Exam	20%
Midterm Exam	10%

- Class attendance is required and critical to success in this course
 - This class moves quickly
- Late assignments will lose 5% each day they are late



What's on Blackboard?

- Class content (including slides)
- Homeworks
- Exams
- Final Project Milestones
- Announcements

Final Project - Overview

Milestone #2: Dimensional Modeling

- Create the logical, Dimensional Models for the data you found in Milestone #1
- You will have ≥ 1 Fact Tables
- You will use the Kimball BUS Matrix to model Dimensions (may be Conformed Dimensions)

Milestone #4: BI Application Foundation

- Map out the Reports/Dashboards/Scorecards you will create to deliver the KPI's you created in Milestone #1
- Properly connect your physical data from Milestone #3 to a BI Application (Tableau)

Milestone #1: Project Planning

- Generate justification for project
- Find ≥ 1 necessary datasets
- Create ≥ 5 KPI's that guide data-driven decisions for your audience

Milestone #3: ETL

- Either manually (with code) or with an application, physically get your data from Milestone #1 into the dimensional models you created in Milestone #2
- Document all steps along the way, and test that your Fact Tables and dimensions connect

Final Project Due

- Use template to submit your Final Project



Course Books

- [Business Intelligence Guidebook: From Data Integration to Analytics](#) by Rick Sherman. Elsevier Science & Technology (Nov 07, 2014). ISBN: 978-0-12-411461-6 Price: 60.00 USD. Also available online at [Baruch Newman Library](#).
- **(Optional)** [The Data Warehouse Toolkit \(3rd. Edition\)](#). by Ralph Kimball and Margy Ross. Publisher: Wiley Edition: 3, Year Published: 2013, Price: 60.00 USD. ISBN: 978-1-118-53080-1
- **(Optional)** [SQL in 10 Minutes](#). by Ben Forta. Publisher: SAMS: 3, Year Published: 2013, Price: 29.00 USD. ISBN: 978-0-672-33607-2



Class Structure

- Each class will look like the following:
 - Briefly review content from last lecture (6:05 - 6:15)
 - Talk about upcoming due dates (6:15 - 6:25)
 - Learn new content (6:25 - 8:30)
 - 10-minute break at 7:30pm
 - *Sometimes, highlight an industry relevant topic (8:00 - 8:30)*
 - Leave time for student questions (8:30 - 9:00)
- **Please bring your laptop to each class.** If you do not have a laptop, contact me and I can help you get a loaner laptop from Baruch.
- I will attempt to make each class as interactive as possible: questions, workshops, mini-projects, etc.



Last notes


- Please let me know if my voice is not loud enough or my microphone is not working
- This course will be interactive, so ask questions (I will also give you room for questions)
- If I don't answer your email within 24 hours, feel free to email again! I want you to succeed and never to be stuck with a question
- Attendance is key to success in this course, I will try to record all classes but the quality will be far better in person
- Please correct me if I mispronounce your name



Syllabus

- Find on Blackboard
 - Familiarize yourself with Blackboard folder structure

10 Minute Break

The background is a solid teal color. In the top-left corner, there are three vertical bars of varying heights, all in a lighter shade of teal. In the bottom-right corner, there are four vertical bars of increasing height from left to right, also in the same lighter shade of teal.

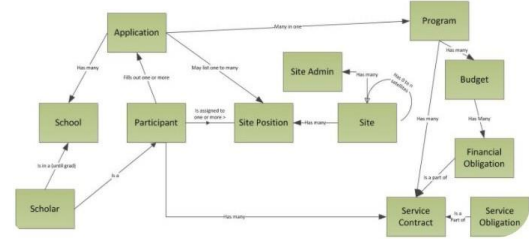
CIS 9340 Review

(The Operational System)

ER Modeling Review

(How confident are you drawing an ER Diagram?)

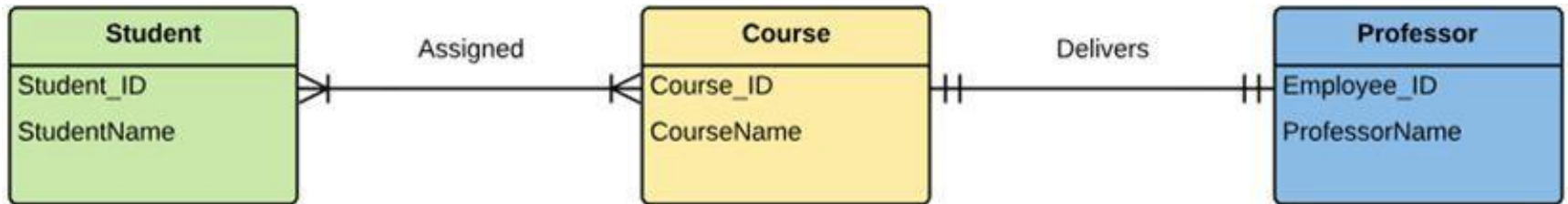
But First, What is Data Modeling?



- Data Modeling is transforming data from source capture into data models. In this course we will use data modeling to transform data into operational data models and dimensional models.
 - Two types of Data Modeling
 - **ER Modeling** - for Operational Data Models
 - **Dimensional Modeling** - for Dimensional Models
- Data Models are specifications of data structures
 - Data Models range in granularity for different audiences
 - Conceptual, Logical, and Physical

What is an ER Diagram?

- What's the purpose of an Entity Relationship Diagram?
 - To visualize and design how all elements/data in a database interact with each other
- Entity Relationship Diagrams are comprised of 3 components:
 - **Entities** - person, place, or thing
 - **Relationships** - links between entities
 - **Attributes** - characteristics of each entity



ER Diagram: Entities

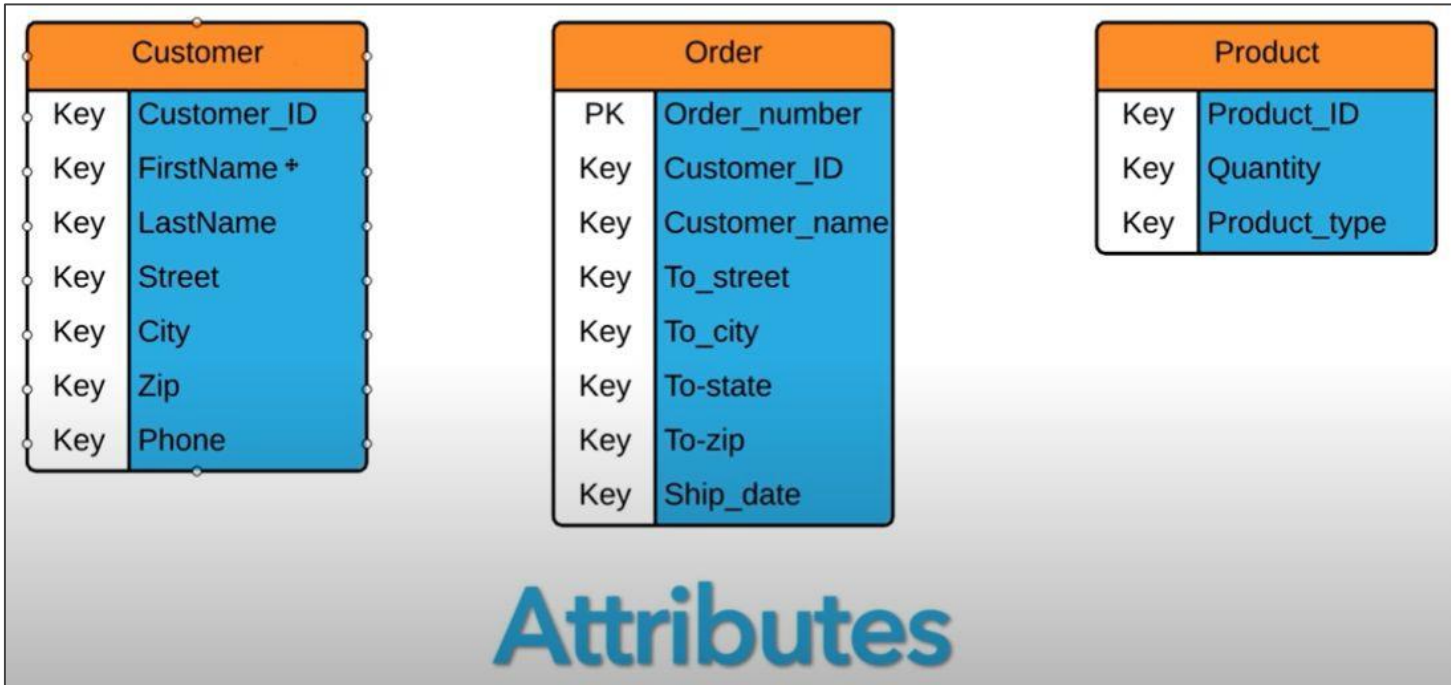
Entities

Customer	
Key	Customer_ID
Key	FirstName
Key	LastName
Key	Street
Key	City
Key	Zip
Key	Phone

Order	
PK	Order_number
Key	Customer_ID
Key	Customer_name
Key	To_street
Key	To_city
Key	To-state
Key	To-zip
Key	Ship_date

Product	
Key	Product_ID
Key	Quantity
Key	Product_type

ER Diagram: Attributes



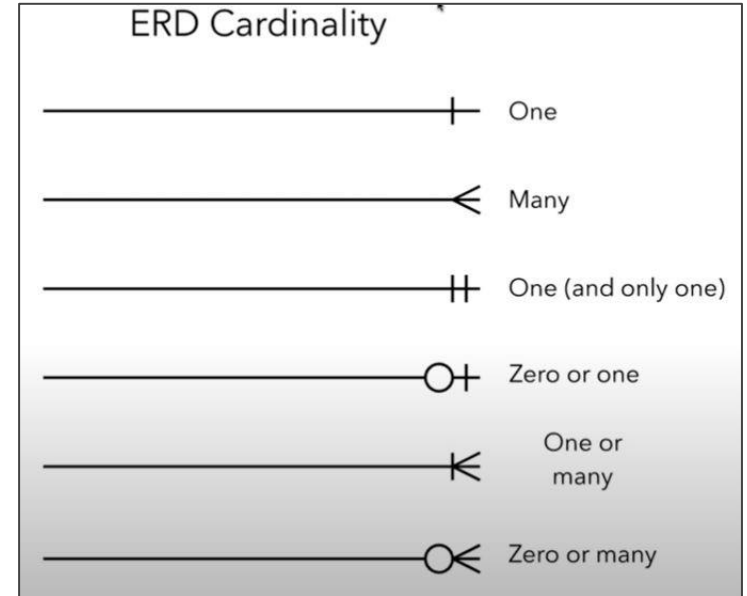
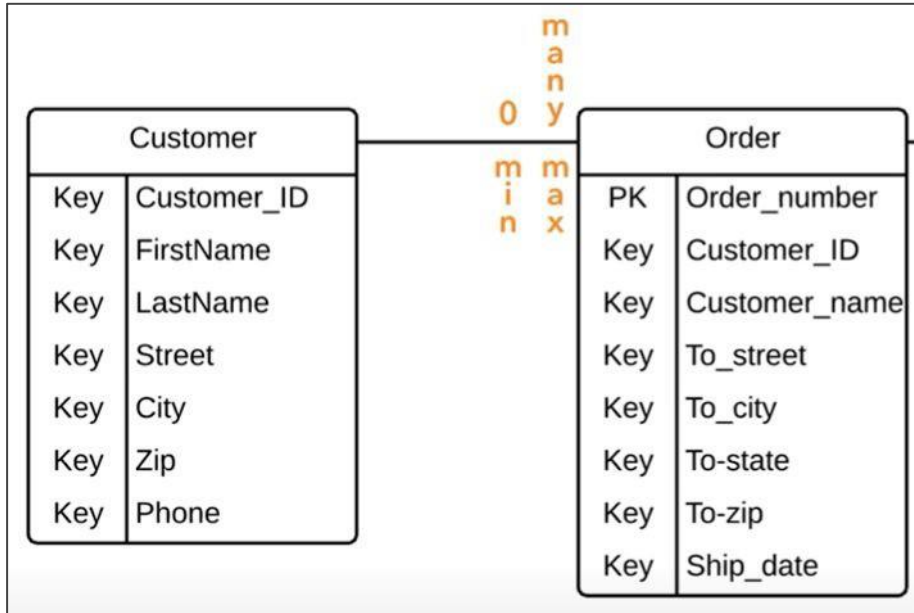
ER Diagram: Entities and Attributes

	Attribute	Attribute	Attribute	Attribute	Attribute
Entity					
Entity					
Entity					
Entity					
Entity					
Entity					
Entity					

DATABASE

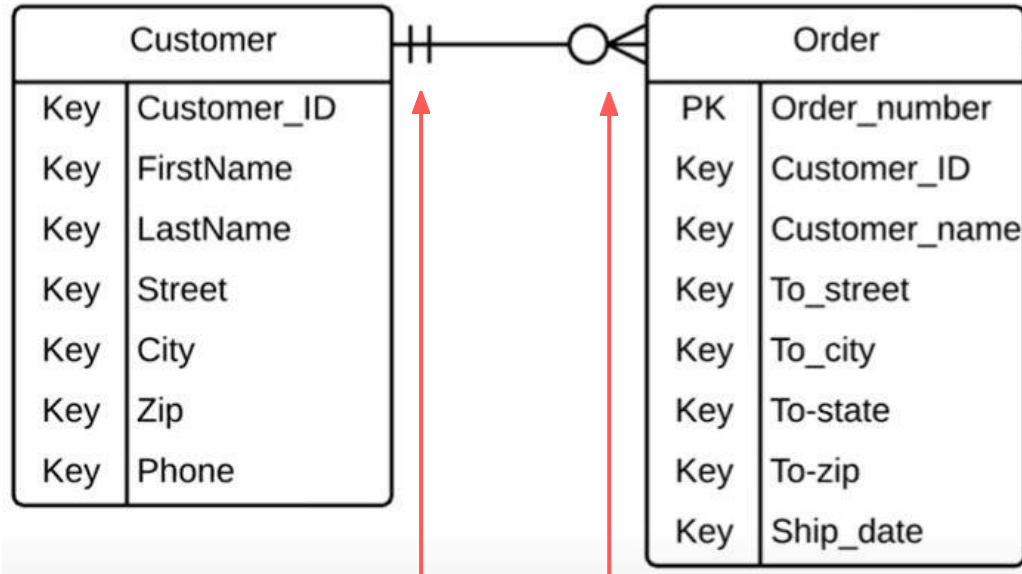
episodes	
episode	varchar
name	varchar
release_date	date
duration_min	integer
external_url	varchar
id	varchar
language	varchar
release_date_precision	carchar
uri	varchar
description	varchar
year_number	integer
month_number	integer
week_number	integer
year_week	integer
year_month	integer

ER Diagram: Relationships



(this is “**Crow’s Foot** notation”)
(relationships are also referred to as “**Cardinality**”)

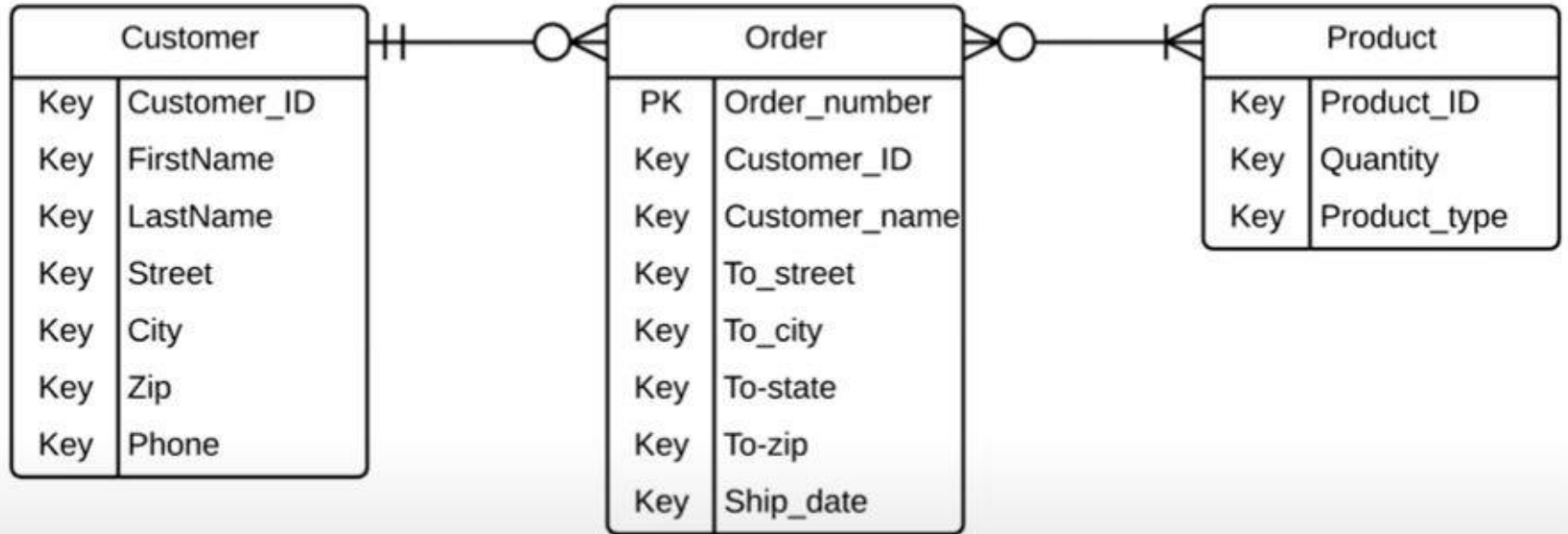
ER Diagram: Relationships (continued)



An order can have **One and Only One** customer

A customer can have **Zero of Many** orders

ER Diagram: Relationships (continued)





ER Diagram: Keys

- **Candidate Keys:** An attribute (or group of attributes) that can uniquely identify each instance of an entity
- **Primary Key:** The attribute chosen to uniquely identify each instance of an entity
- **Alternate Keys:** Attributes that can be a primary key, but are not chosen for the primary key for their entity
- **Foreign Key:** A primary key of a parent entity. One entity can have multiple Foreign Keys.
- **Composite Key:** A key of two or more attribute that uniquely identifies the row (i.e., OrderDate and CustomerID)

ER Diagram: Primary Keys

PK Requirements:

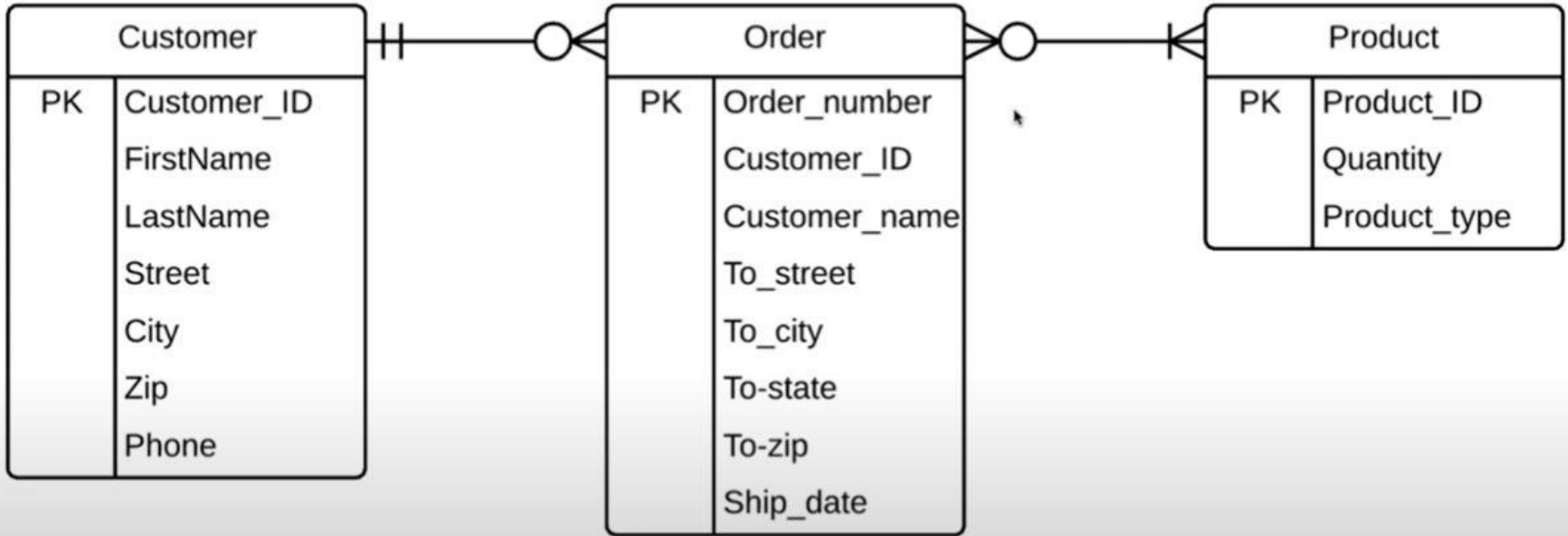
1. Unique
2. Never changing
3. Never null

Which attribute would you pick for a Primary Key of this Customers table?

Why don't other attributes work?

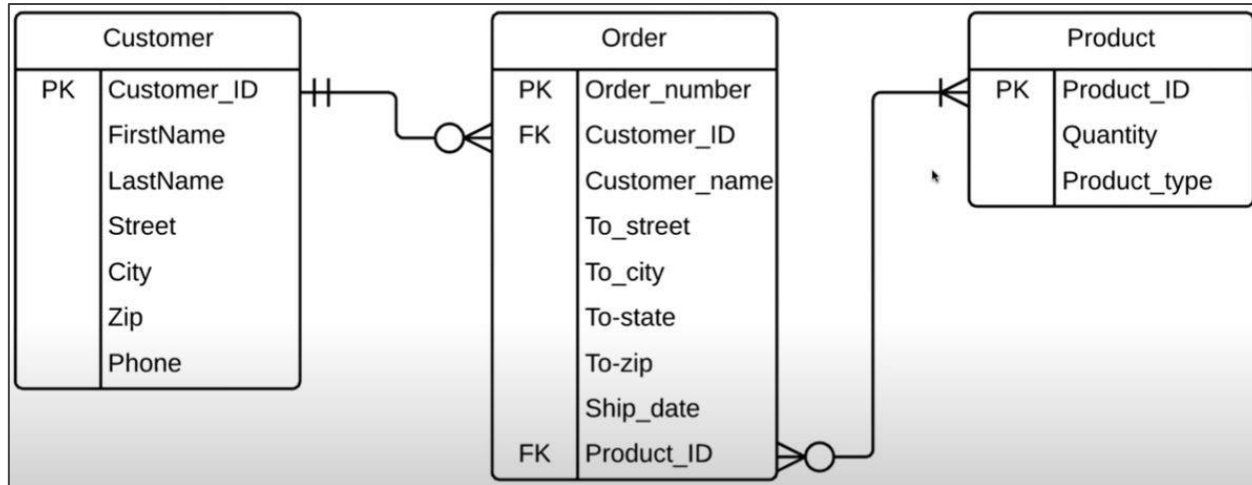
1	Customer_ID	FirstName	LastName	Street	City	Zip	Phone
2	30001	Jacob	Henderson	32 Myers St.	Phoenix	24635	895-698-4314
3	30002	Noelle	Sanderson	24 Hill Drive	Banning	72025	597-502-3005
4	30003	Spencer	Martin	8406 St Margarets St.	Langhorne	90025	226-457-1686
5	30004	Solange	Carr	8272 Durham St.	Beckley	93382	478-870-3240
6	30005	Tony	Bell	70 Hawthorne Street	Meriden	57812	935-295-0925
7	30006	Laurel	Hudson	722 Newcastle Ave.	West Fargo	11814	429-793-0704
8	30007	Ken	Turner	8579 Vernon Rd.	Menasha	84699	249-928-4530
9	30008	Charlotte	Burgess	444 Central Lane	Chicago	44476	997-192-5311
10	30009	Ben	Ellison	9533 S. Purple Finch Lane	Easton	54985	100-576-0463
11	30010	Danika	Marshall	44 Jones Road	Manchester	16685	610-934-2799
12	30011	Linda	McGrath	7249 N. Bow Ridge St.	Ft Mitchell	23358	903-296-6663
13	30012	Iris	Edmunds	7135 North Rocky River Court	Yorktown	55720	728-849-9825
14	30013	Chandra	Parsons	847 Tanglewood Dr.	Calhoun	27759	821-271-9463
15	30014	Ranee	Peters	696 Fawn Court	Albany	97083	614-522-4822
16	30015	Steven	Langdon	64 Pennington Ave.	Jacksonville	33490	545-041-1643
17	30016	John	Smith	7411 Shirley Street	Springfield	41437	522-287-2538
18	30017	Ben	Chapman	6 James Ave.	Hopkinsville	30476	172-245-1141
19	30018	Jeremy	Nash	76 Strawberry Court	Billerica	70728	111-267-2814
20	30019	Rhett	Buckland	243 Mayflower St.	Watertown	97924	147-612-1745

ER Diagram: Primary Keys (continued)



ER Diagram: Foreign Keys

- **Foreign Keys:** A reference to a Primary Key of a different entity
 - Foreign Keys do not have to be unique, they can repeat
 - You may have multiple Foreign Keys in one entity



ER Diagram: Composite Primary Keys

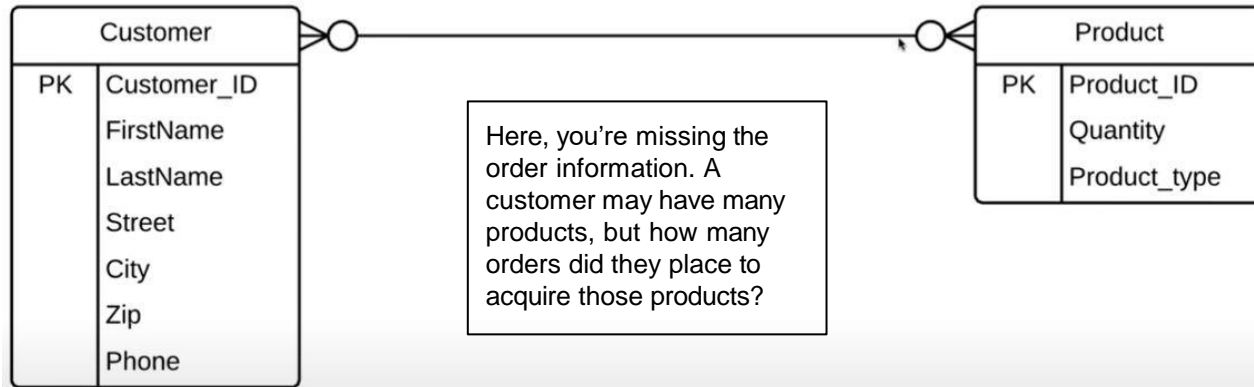
- **Composite Key:** A Primary Key of two or more attribute that uniquely identifies the row
 - Use as few attributes as possible, and don't use if not needed

Shipment	
PK,FK	Product_ID
PK,FK	Order_number
	ChargeCardTime
	PackingTime
	ShipOrderDate

Composite Primary Key			Shipment Table		
1	Product_ID	Order_number	ChargeCardTime	PackingTime	ShipOrderDate
13	49225	252349915	6/1/2017 9:13:34	6/2/2017 10:14:46	6/3/2017 11:15:52
14	40807	252349916	6/1/2017 9:14:16	6/2/2017 10:15:02	6/3/2017 11:16:03
15	76342	252349917	6/1/2017 9:14:01	6/2/2017 10:15:26	6/3/2017 11:16:13
16	96893	252349918	6/1/2017 9:14:21	6/2/2017 10:15:39	6/3/2017 11:16:19
17	69246	252349919	6/1/2017 9:14:34	6/2/2017 10:15:41	6/3/2017 11:16:47
18	69253	252349919	6/1/2017 9:14:34	6/2/2017 10:15:45	6/3/2017 11:16:47
19	99002	252349920	6/1/2017 9:15:07	6/2/2017 10:16:07	6/3/2017 11:17:11

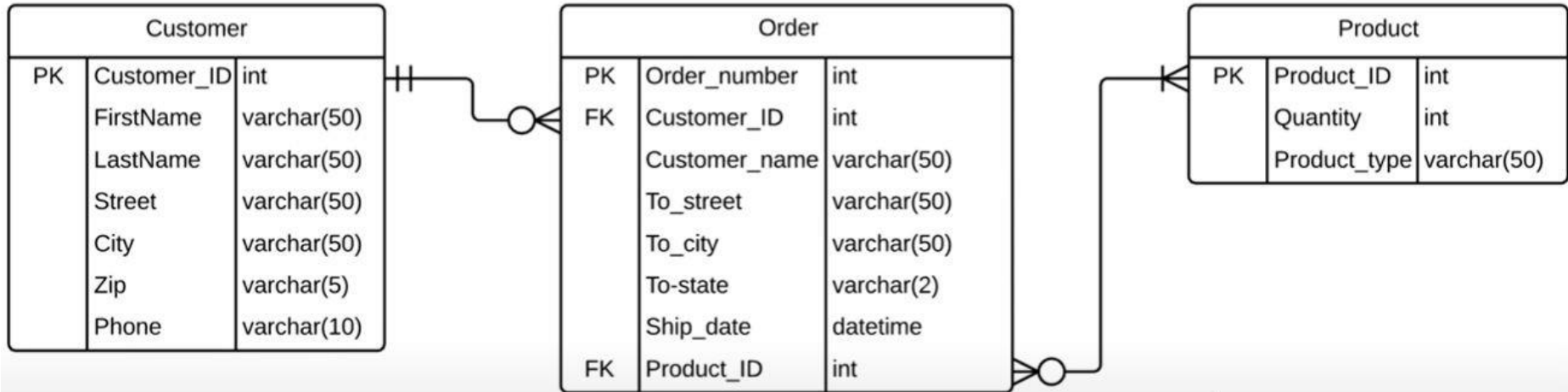
ER Diagram: Bridge Table

- **Bridge Table:** A table put between two entities to capture information you are missing in the relationship
 - Look to create a Bridge Table is when you have a Many-to-Many relationship



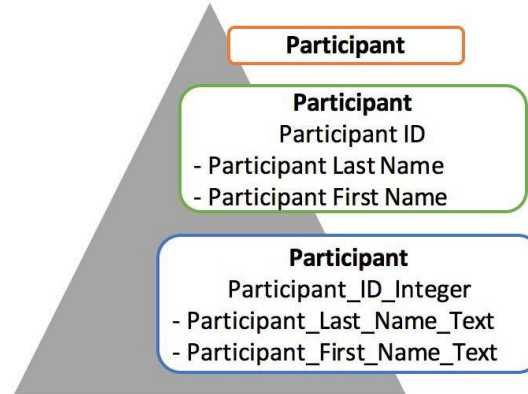
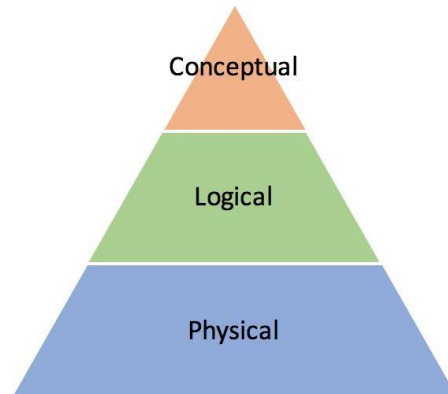
ER Diagram: Add Data Types (Physical)

- To turn your ER Diagram from a Logical Model to a Physical Model, add data types



Three Types of Data Models

- **Conceptual:** business requirements, business users
- **Logical:** architecture requirements, business analysts
- **Physical:** developer requirements, database developers



Are ER Models
Conceptual,
Logical, or
Physical?

Exercise: Draw your own ER Model

(20 minutes)

1. Go to [Lucidchart.com](https://lucidchart.com), create an account if you do not already have one
2. Click on “Documents”
3. Click “New -> Lucidchart -> Create from Template”
4. Choose “Database diagrams” then “Database ER diagram (crow’s foot)”
5. Create an ER Diagram with at least 3 entities

Example Topics:

- University - Classes, Students, Grades, Attendance
- Ski Resort - Trails, Lifts, Skiers
- Literature - Writers, Books, Articles, Readers
- Netflix - Movies, Shows, Actors, Genres
- Google “ER Model examples”

Normalization Review

(From CIS 9340)



What is Normalization?

- Normalization is the process to reduce redundancy in a data model.
 - **Upside:** In Third Normal Form (3NF), each attribute that belongs to an entity will have a unique position within the Data Model
 - **Downside:** If Normalization is very strictly enforced, or beyond 3NF, database performance (speed) is compromised due to increased number of joins. The higher the number (4NF, 5NF, etc) the stricter it is.

First Normal Form (1NF)

* Pictures from 1KeyData.com

- **1NF:** No repeating groups

TABLE_PRODUCT

Product ID	Color	Price
1	<u>red, green</u>	15.99
2	yellow	23.99
3	green	17.50
4	yellow, blue	9.99
5	red	29.99

1NF



TABLE_PRODUCT_PRICE

Product ID	Price
1	15.99
2	23.99
3	17.50
4	9.99
5	29.99

TABLE_PRODUCT_COLOR

Product ID	Color
1	red
1	green
2	yellow
3	green
4	yellow
4	blue
5	red

Second Normal Form (2NF)

* Pictures from 1KeyData.com

- **2NF:** No partial-key dependencies

TABLE_PURCHASE_DETAIL

Customer ID	Store ID	Purchase Location
1	1	<u>Los Angeles</u>
1	3	San Francisco
2	1	<u>Los Angeles</u>
3	2	New York
4	3	San Francisco

2NF →

TABLE_PURCHASE

Customer ID	Store ID
1	1
1	3
2	1
3	2
4	3

TABLE_STORE

Store ID	Purchase Location
1	Los Angeles
2	New York
3	San Francisco

Third Normal Form (3NF)

* Pictures from 1KeyData.com

- **3NF:** No non-key interdependencies

TABLE_BOOK_DETAIL

Book ID	Genre ID	Genre Type	Price
1	1	Gardening	25.99
2	2	Sports	14.99
3	1	Gardening	10.00
4	3	Travel	12.99
5	2	Sports	17.99

3NF



TABLE_BOOK

Book ID	Genre ID	Price
1	1	25.99
2	2	14.99
3	1	10.00
4	3	12.99
5	2	17.99

TABLE_GENRE

Genre ID	Genre Type
1	Gardening
2	Sports
3	Travel



Normalization Summary

Normalization Levels	Normalizing an Entity
First normal form (1NF)	Eliminate repeating groups. Make a separate table for each set of attributes (in essence, this is creating an entity). Identify a primary key for each table. If you cannot define a primary key, then you have not split up the tables into the sets of related attributes creating an entity, and you need to repeat this step
Second normal form (2NF)	Eliminate redundant data stored in different entities. If an attribute depends on anything other than the primary key (could be a compound key) then remove it as a separate table
Third normal form (3NF)	Eliminate non-key interdependencies. If you have defined the primary and the keys within that, then all the attributes in that entity need to be related to that key. For example, if you have customer or product, you can only have attributes that are related to the customer or product within the entity. Otherwise, remove them and put them into a separate table, as they are most likely separate entities. With these steps completed, you have defined a 3NF schema



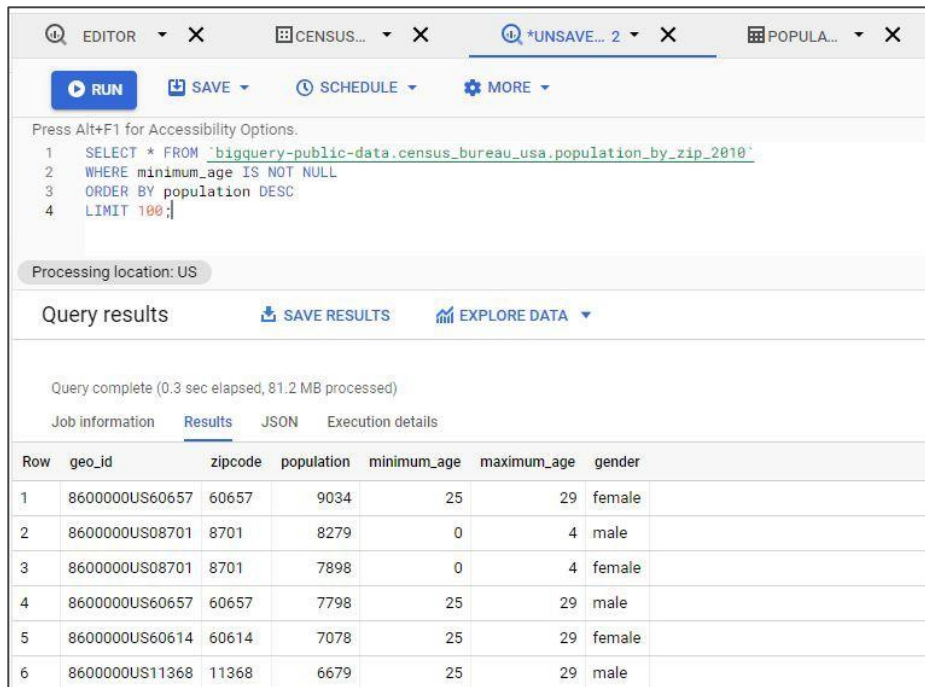
Normalization Details

- **Business Intelligence Guidebook**, pages 189-192

The background is a solid teal color. In the top-left corner, there are three vertical bars of varying heights, all in a lighter shade of teal. In the bottom-right corner, there are four vertical bars of increasing height from left to right, also in the same lighter shade of teal.

Intro to BigQuery: [link](#)

BigQuery Example



The screenshot displays the Google Cloud BigQuery console. At the top, there are tabs for 'EDITOR', 'CENSUS...', '*UNSAVE... 2', and 'POPULA...'. Below the tabs is a toolbar with buttons for 'RUN', 'SAVE', 'SCHEDULE', and 'MORE'. The main area contains a SQL query:

```
1 SELECT * FROM `bigquery-public-data.census_bureau_usa.population_by_zip_2010`  
2 WHERE minimum_age IS NOT NULL  
3 ORDER BY population DESC  
4 LIMIT 100;
```

Below the query, it indicates 'Processing location: US'. Underneath, there are links for 'Query results', 'SAVE RESULTS', and 'EXPLORE DATA'. The status bar shows 'Query complete (0.3 sec elapsed, 81.2 MB processed)'. Below this, there are tabs for 'Job information', 'Results', 'JSON', and 'Execution details'. The 'Results' tab is active, displaying a table with 6 rows and 7 columns: 'Row', 'geo_id', 'zipcode', 'population', 'minimum_age', 'maximum_age', and 'gender'.

Row	geo_id	zipcode	population	minimum_age	maximum_age	gender
1	8600000US60657	60657	9034	25	29	female
2	8600000US08701	8701	8279	0	4	male
3	8600000US08701	8701	7898	0	4	female
4	8600000US60657	60657	7798	25	29	male
5	8600000US60614	60614	7078	25	29	female
6	8600000US11368	11368	6679	25	29	male

The background is a solid teal color. In the top-left corner, there are three vertical bars of varying heights, all in a lighter shade of teal. In the bottom-right corner, there are four vertical bars of increasing height from left to right, also in the same lighter shade of teal.

Intro to NYC Open Data: [link](#)

Homework:

1. Get started with Google BigQuery: [link](#)
2. Review Syllabus, explore Blackboard
3. Reading, BIG Chapter 8
4. Complete “Before Class 1” checklist
5. Discussion Board - Introduction
 - a. Feel free to reply to others