Fall 2017 - W205 – Storing and Retrieving Data
Week 4 Live Class Session Agenda
Kevin R. Crook

- Course being split into two courses, targeting Spring 2018, but probably Summer 2018

- Schedule
  - Lab 5 – due Tuesday, 9/26/2017 at 11:59 pm
    - In ISVC, please submit under Week 7, Lab 5
  - Asynchronous for next week
    - Unit 4 – Data Lakes: Storage and Maintenance
  - Exercise 1
    - Distributed today
    - Due Tuesday 10/31/2017 at 11:59 pm

- SQL
  - Lab 5 – PostgeSQL
    - In addition, you may want to go through the tutorial, a link is on my website in the videos section for the lab
  - Lab 3 – HiveSQL
    - In addition, you may want to go through the tutorial, a link is on my website in the videos section for the lab
  - On my website, book recommendations

- Labs
  - Labs are a "jumpstart" to get you up and running with a technology so you can learn more about that technology
  - Some students complain that they are just copying and pasting commands in the labs and not understanding what the command mean – this is a student's choice – 100% within the student's control

- Grading
  - Labs
    - Easy, not very subjective, most students get high marks if they complete them, lots of 100%'s
    - Instructor can help you with the labs, and give you hints to answer the questions, but cannot answer them for you

- Exercise 1, 2, and the Project are graded differently
    - Much harder than the labs
    - 90% - means that you met the minimum requirements - objective
    - Above 90 % - means excellence above and beyond the minimum requirements - subjective
    - Above 94 % - rare
    - Instructor help
        - Yes
            - Technical issues – basically the parts covered in the videos on my website
        - No – these are research items that are part of the exercise
            - Questions about the data, what fields mean, what values are good, what data is missing, what calculations are correct, etc.
            - Questions about how to structure the data
            - Questions about analytics other than high level
            - Asking for a check list of items that you need to look at for your analytics – look at and analyze as much as possible!
        - No pre-grading
            - Please don't send me a copy in email and ask me to pre-grade it – not fair to other students for some students to get a pre-grading and a chance to correct their submission
    - Think of it this way – if a company hired you as a high dollar consultant for data science, could you hand them your exercise or project one day and a $25k bill the next day and be comfortable that they would pay it?

- Asynchronous
    - Today in class
        - we will do a break out soft skills exercise
        - Unit 7 – Querying Data
        - Unit 3 – Structure and Organization
    - Next week in class
        - We will do an instructor led exercise on DAGs, SQL, and SQL execution plans
        - Unit 4 – Data Lakes: Storage and Maintenance

(next page)

Break Out Exercise
"Soft Skills" – Applying Theory from the Asynchronous to Real World Examples
Procedural Languages, Declarative Languages, Mixing the Two
Sampling: Uniform Sample, Stratified Sample
ACID versus BASE
Reading an Entity-Relationship Diagram

(All Groups)
Procedural Languages, Declarative Languages, Mixing the Two

Give a real world example of when a declarative language (such as SQL) would be used instead of a procedural language (such as Python).

Give a real world example where you use SQL for something, and you hit a point where you need a procedural language (such as a loop).

Give a real world example of a problem that can best be solved using Python with SQL. What parts would you use SQL for? What parts would you use Python for?

(All Groups)
Sampling: Uniform Sample, Stratified Sample

Give a real world example where sampling was improper and resulting analytics were wrong.

Give a real world example of a statistical experiment and how you would ensure the sampling is proper.

(All Groups)
ACID versus BASE

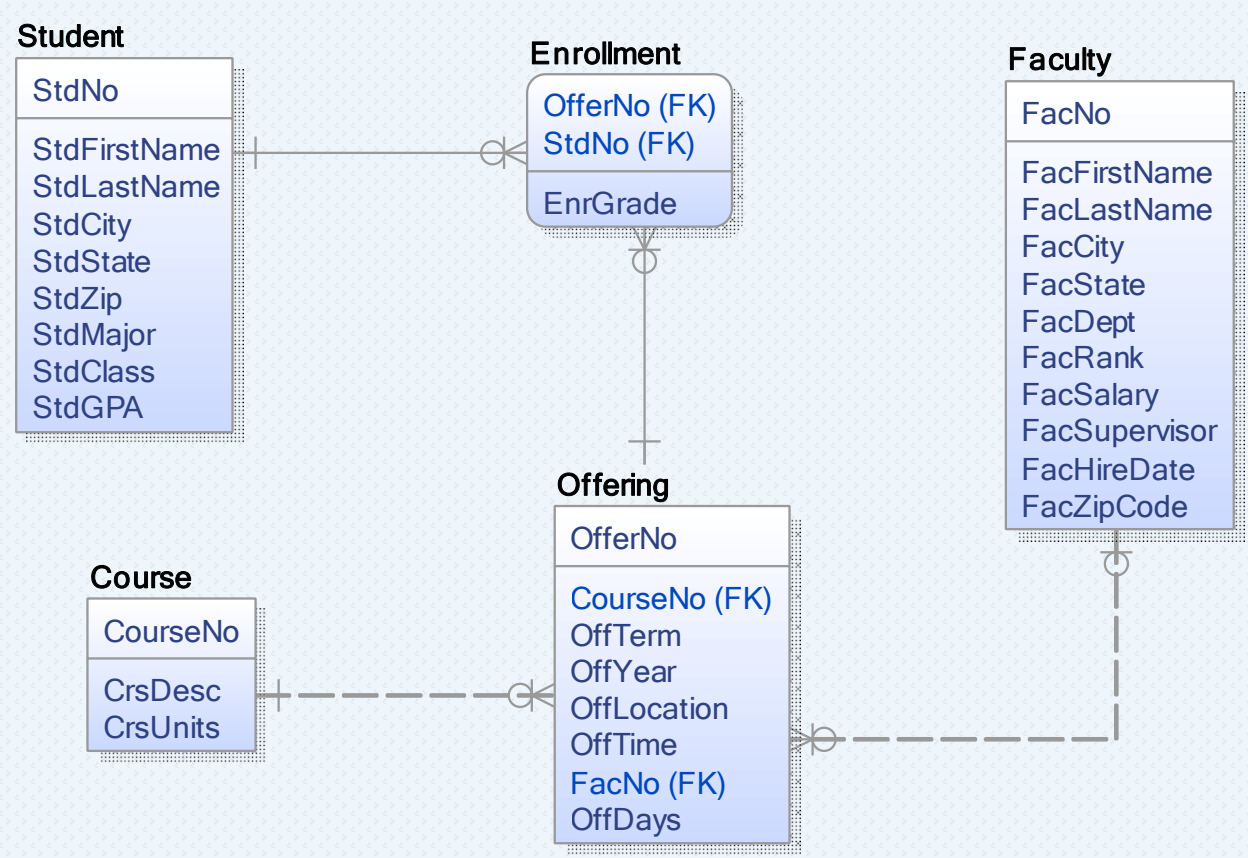
Give a real world example of a database transaction where it needs to be ACID.

Give a real world example of a database transaction where we can relax it to BASE.

(All Groups)
Reading an Entity-Relationship Diagram

Each group will have an ERD to read. Describe the entities, the primary keys, relationships, foreign keys, and any diagram specific questions I have put.

(Group 1)



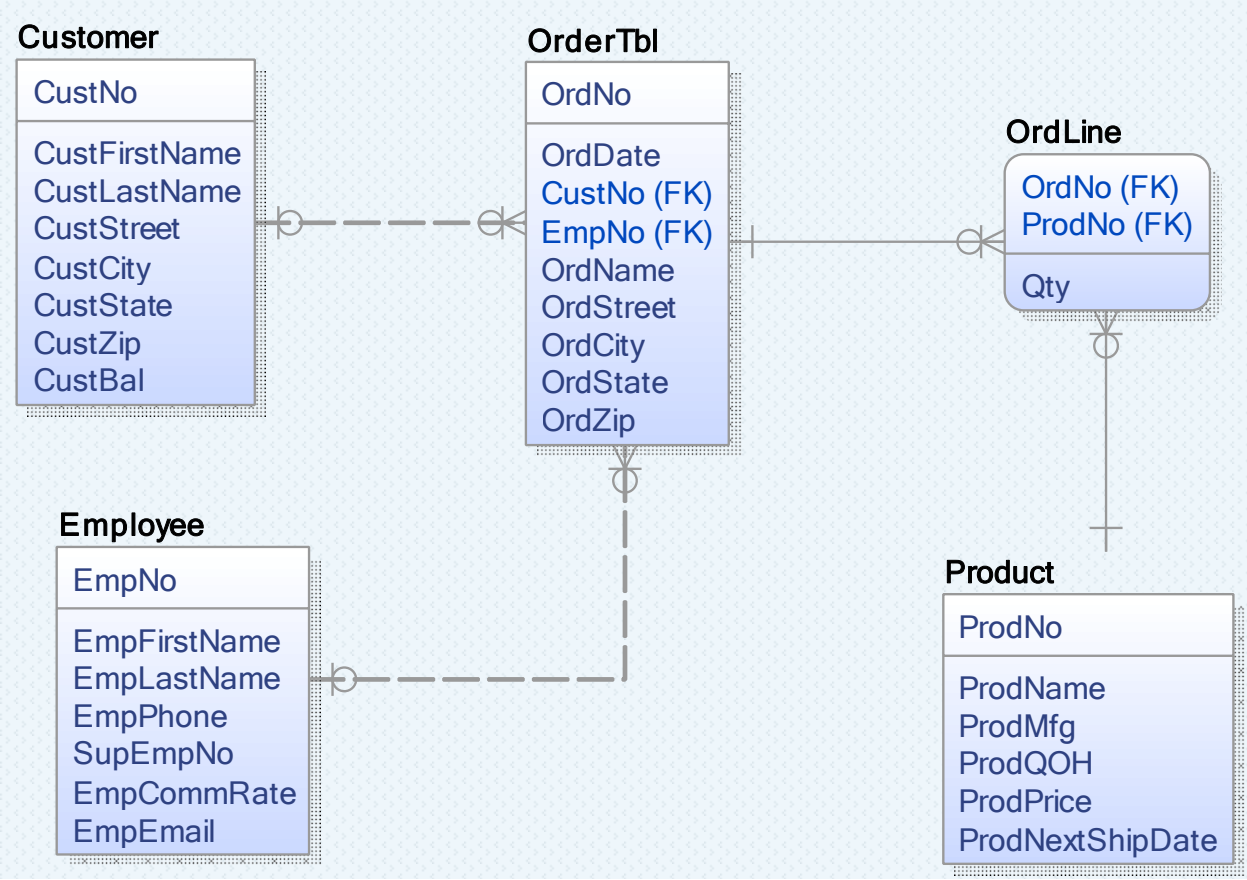
Extra questions:

Can there be a course offering without a faculty? If not, how would we change the ERD to make this possible?

Can there be faculty without courses? If not, how would we change the ERD to make this possible?

Can there be a course offering with more than one faculty? If not, how would we change the ERD to make this possible?

(Group 2)



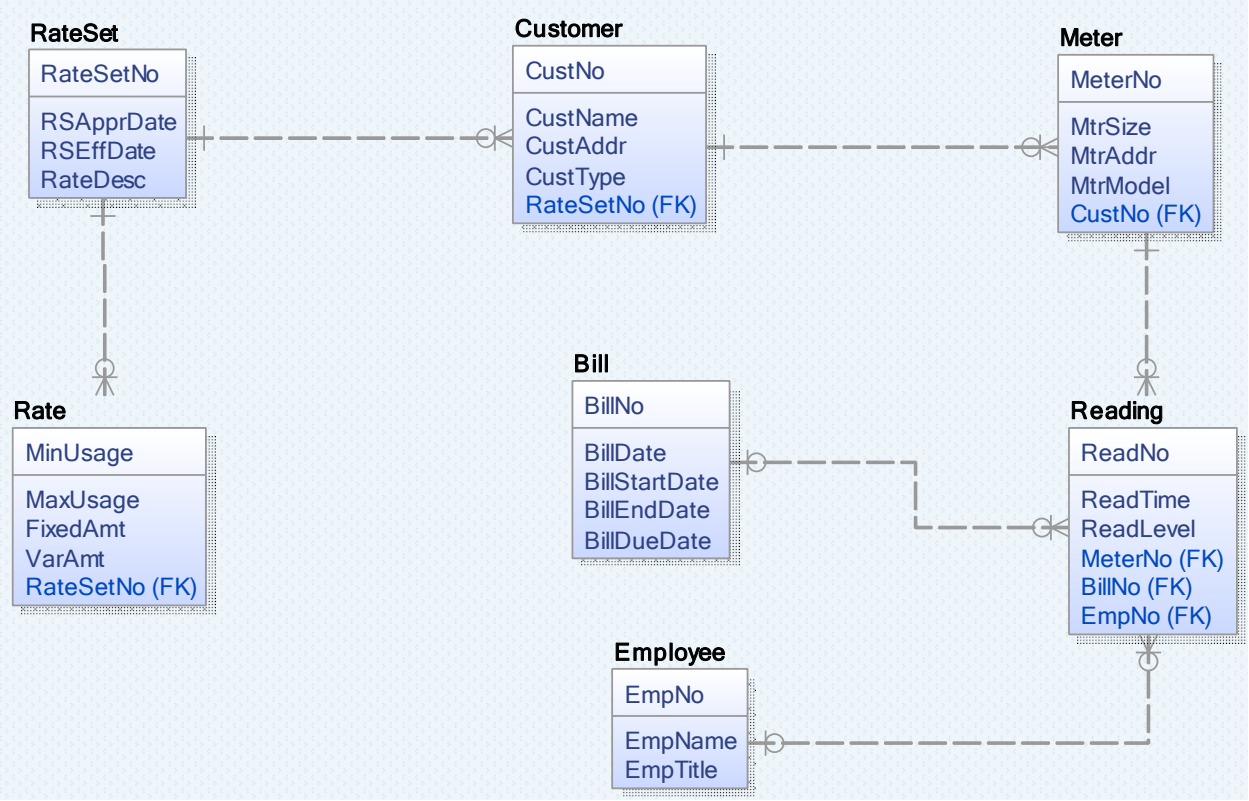
Extra Questions:

Can there be an order without an employee? If not, how would we change the ERD to make this possible?

Can we have an order without a customer? If not, how would we change the ERD to make this possible?

Can we have an order line with more than one product? If not, how would we change the ERD to make this possible?

(Group 3)



Is the primary key for Rate proper? If not, how would we change the ERD to fix it?

Can a meter belong to more than one customer? If not, how would we change the ERD to fix it?

Can a meter be read by more than one employee? If not, how would we change the ERD to fix it?

Does a meter have to be read by an employee? If not, how would we change the ERD to fix it?