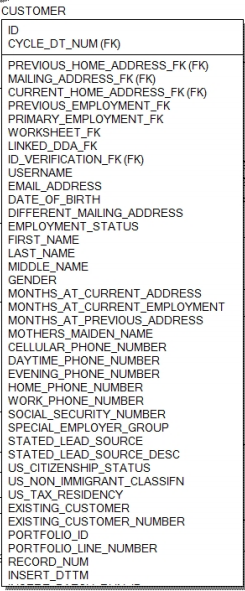
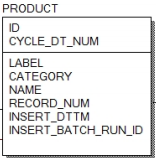
**SQL Exercise:**



**Key:**

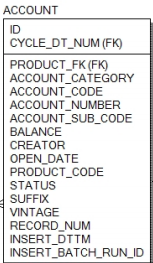
application.product\_fk = product.id

0..1

1..1

**Key:**

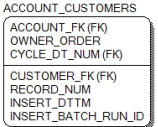
account.product\_fk = product.id



**Key:**

customer.application\_id = application.id

1..N



**Key:**

customer.id = account\_customers.customer\_fk

1..1

**Key:**

account\_customers.account\_fk = account.id

0..N

**Background:**

* This is an example schema used by banks to track online applications for accounts.
  + The **Customer** table represents all of the customers that have applied for an account through the online application process.
  + The **Application** table represents an actual electronic application submitted by a customer. There may be multiple apps per customer
  + The **Product** represents the product that was selected during the application process. There may be multiple products per application.
  + The **Account** describes the account that was opened during the application process
  + 1..N means a one-to-many relationship, 0..N means zero-to-many, and 1..1 means one-to-one, 0..1 means zero or one row. Arrows denote the join / relationship direction.
  + State any assumptions you’re making about the data as part of the exercise. **It’s OK to make assumptions about field definitions based on the field name**.
  + This exercise tests your ability to deal with **new or uncertain data sets**. There may be multiple correct answers – please select the answer that highlights a succinct, well-thought out approach. Use SQL to express the correct answer when asked – **exact syntax is less important than the thought process behind the SQL**.
  + It’s in your best interest to do your own work and be ready to defend your approach. **Don’t cheat.**

**Consider the following questions:**

* The marketing team is considering new creative material that might have broader appeal to customers. They want to know the characteristics of customers who have already *applied* and *successfully* opened up accounts. At a high-level (without SQL), what table(s) and field(s) would you analyze to provide this information?
  + I will look at all the tables given above. I will classify the characteristics into two categories – Core & Extended
  + Core characteristics
    - From customer table, I will look at the following fields (the column name or field name also can be noted as features/characteristics)
      1. Date\_of\_Birth
      2. Employment\_status
      3. Zip code (mailing address & current address) extraction
      4. Gender
      5. Special\_employer\_group
      6. US\_Citizenship\_Status
      7. US\_Non\_Immigrant\_Classifin
  + Extended characteristics
    - From Application Table
      1. Will look at the number of applications for each successful customer
    - From Accounts Tables
      1. Will look at the number of accounts for each successful customer
    - From Products Table
      1. Will look at the how many products for each successful customer
  + As the marketing team is looking to broader appeal (not targeting or micro targeted marketing)
    - Should consider the characteristics in the following order
      1. US Citizenship status
      2. Employment status
      3. Gender
      4. Date of Birth
      5. Zipcode
      6. Special employer group
      7. US non immigrant classifin
* You have been asked to do a break-down of application statuses (using the *Status* field in the application table) by daytime and evening phone numbers of customers whose application status is ‘pending’. What SQL would you write to answer this question?

Select application.status, customer.first\_name, customer.daytime\_phone\_number, customer.evening\_phone\_number from application, customer where customer.application\_id = application.id and application\_status=’pending’

Please note that customer.application\_id not present in the customer table but given in the ER diagram.

* The digital sales team is considering the value of high-balance accounts (any account with an opening balance greater than $2K) and believes there may be a relationship between *opening balance* and *time of day when the application was submitted*.
  + Please note: the Balance field in the Account table contains the opening balance. The submission time in the application table has when the app was submitted.
    - What SQL would you write to show the relationship between opening balance and submission time?

Select customer.id,account.balance, application.submission\_time from customer, application, account, account\_customers where customer.application\_id = application.id and customer.id = account\_customers.customer\_fk and account\_customers.account\_fk = account.id and account.balance > 2000.00

Please note that customer.application\_id not present in the customer table but given in the ER diagram.

* + - What validation tests would you perform to determine if this is a *sound hypothesis*? Please be specific.
      1. Bucket the 24 hours in to 24 (1 bucket for every one hour) or 48 (1 bucket for every 30 minutes ) different slots/buckets (assuming application submission process is available all day (24 hours))
      2. Populate the buckets with number of counts where the opening balance is great than 2000.00
      3. Descending Order the buckets and the top bucket and associated time where opening balance had greater than 2000.00 and check the variances. If you plot this data and if we see a bell curve and the hypothesis is valid, if you see normal distribution and the hypothesis is wrong as there is no much variance among the buckets
    - Assume the hypothesis about opening balance and time-of-submission is proven false. What other fields would you suggest or investigate, if any, which might yield better *causative* relationships for the opening account balance?
      1. We would alternatively investigate on the date of birth, employment status, zip code (mailing address or current address) and gender independently or in combination