# Solutions for the Module 3 Assignment

1. The dimensions in the problem are somewhat difficult because some dimensions should be combined from data sources. Item and customer combine parts of both data sources. The calendar dimension is a standard, hierarchical dimension. Email can be parsed to be hierarchical as part of the customer dimension.
   - Franchise
     - FranchId: retail database only
     - FranchRegion retail database only
     - FranchPostalCode: retail database only
     - FranchModelType: retail database only
     - For the special events worksheet, selected franchises maintain spreadsheets so the franchise will be derived from the spreadsheet submission.
   - Calendar
     - Date columns in the retail database (SalesDate, ServPurchDate, and MmbrDate) and spreadsheet (EventDate); hierarchical (year → month → day)
   - Item: combines Merchandise, Service, and Events
     - MerchId (Merchandise table) | ServId (ServiceCategory table)
     - MerchName (Merchandise table) | ServCatName (retail database) | Event Name (spreadsheet)
     - MerchType (Merchandise table) | Event Type Code (spreadsheet)
   - Customer: combines members and corporate customers
     - MmbrId (retail database) | Corporate Customer Id (spreadsheet)
     - MmbrName (retail database) | Corporate Customer Name (spreadsheet but must be parsed)
     - Corporate Customer Location (spreadsheet): must be parsed
     - MmbrEmail: retail database; hierarchical (top level domain → second level domain → local part)
     - MmbrZip: retail database only
     - MemTypeId: retail database only

Self-evaluation guidelines

- Award 10 points if the solution has a list of dimensions with at least 1 dimension having a hierarchy.
- Deduct 3 points if no hierarchical dimensions are specified or hierarchies are specified incorrectly.

2. The measures come from several tables in the retail database and special events spreadsheet. Measures from related tables are important to associate with the measures from the PurchLine table and Supply Purchases spreadsheet.
   - Qty (Contains table); additive measure
   - MerchPrice (Merchandise table); non additive measure
   - ServCatPrice (ServiceCategory table) | Amount (spreadsheet); non additive if considered a price; additive if considered revenue

Self-evaluation guidelines

- Award 10 points if the solution has a list of dimensions with at least 1 dimension having a hierarchy.
- Deduct 3 points if no hierarchical dimensions are specified or hierarchies are specified incorrectly.

3. The most detailed grain is the combination of individual customer, product or service, and date. The franchise is not a direct factor in the grain unless only the average customers per franchise are used in the grain calculation. The grain should include service purchases, event occurrences, and merchandise sales.
    - Franchises
        - 350 franchises for merchandise sales
        - 200 franchises for special events
        - Assume that franchises for special events are already included in the franchises with merchandise sales
    - Items
        - Merchandise: 500
        - Service categories: 20
        - Event types: 1
        - Total types of goods/services for sale: 521
    - Members and customers
        - Members: 50,000 members in the retail database
        - Corporate customers: 150 customers * 200 franchises (30,000)
        - Total members and customers: 80,000 assuming no overlap
    - Fact table size
        - Merchandise purchases: rows in the Contains table (450,000/year)
        - Service purchases: rows in the ServicePurchase table (100,000/year)
        - Special events: worksheet rows (300 events * 200 franchises = 60,000 events per year)
        - Total rows: 610,000 rows per year
    - Sparsity estimate
        - Franchise is not required to compute the fact table size and sparsity. Customer and date determine the franchise so franchise can be ignored in the calculations.
        - 1 - ( fact table size / product of dimensions )
        - (1 – ( 610,000 / (521*365*80,000) ) = 0.999959903
        - The data cube has mostly missing cells with less than 1% of cells with non zero values. More than 99% of cells are empty.

Self-evaluation guidelines

- Award 10 points if the solution has cardinality or size estimates for each dimension (customer, item, and date) and the fact table. The franchise is not a direct factor in the grain unless only the average customers per franchise are used in the grain calculation. The solution should show a sparsity estimate using the formula 1 – (fact table size / product of dimensions). The fact table cardinality should be derived from the rows in the Contains table, ServicePurchase table, and special events worksheet with total rows about

610,000. For dimensions, the cardinality estimates should be 521 (items), 80,000 (customers), and 365 (dates). For 10 points, a student does not need to show the same values for the cardinalities of fact and dimension tables.

- Deduct 3 points if the sparsity formula is not estimated or the wrong formula appears to have been used.
- Deduct 5 points if more than 2 elements are missing in the solution such as missing sparsity estimate and missing cardinalities for dimension or fact tables.

4. The star schema should support the dimensions and measures specified in problems 1 and 2. Franchise was related directly to RevFact instead of customer. All revenue is associated with a franchise. Corporate customers can be associated with multiple franchises so the relationship to RevFact was used. RevFact contains merchandise sales, service purchases, and event occurrences. An alternative design is to split into multiple fact tables. The data warehouse design will be more complex with a constellation schema. The ItemType column in the Item table supports differentiation between merchandise, service, and event revenue so the design with one fact table is preferred.

   The flatten and merge transformations apply to the tables of the Retail Fitness database. The *RevFact* table involves flatten and merge transformations. A flatten transformation applies to the *Sale* and *Contains* tables of the Retail Fitness database. The *SalesDate* and *MmbrId* columns group products purchased together. After the flatten transformation, a merge transformation applies to the flattened table with the *ServicePurchase* table. The *Item* table involves a merge of *Merchandise* and *ServiceCategory* tables.

   The merge transformation combines transformed tables of the Retail Fitness database and special events spreadsheet. The *Member* table and spreadsheet columns (Corporate Customer Name) merge into the *Customer* table. The *RevFact* table combines transformed tables (flatten and merge) of the Retail Fitness database and columns of the Special Events Worksheet (Event Date and Amount). Event columns (Event Type Code and Event Name) in the Special Events Worksheet merge with the transformed *Item* table of the Retail Fitness database.

   Transformation, Source Objects, Result Object, Comments

| **Transformation** | **Source Objects** | **Result Object** | **Comments** |
|---|---|---|---|
| Flatten | Sale, Contains | RevFact | New primary key column |
| Merge | RevFact, ServPurchase | RevFact | ServPuchDate combines with SalesDate |
| Merge | ServiceCategory, Merchandise | Item | New primary key column |
| Merge | Member, Special Events Worksheet (Corporate Customer Name) | Customer | New primary key |

| | | | |
|---|---|---|---|
| Merge | RevFact, Special Events Worksheet (Event Date and Amount) | RevFact | Assign new primary key values for special events |
| Merge | Item, Special Events Worksheet (Event Type Code and Event Name) | Item | Assign new primary key values for special events |

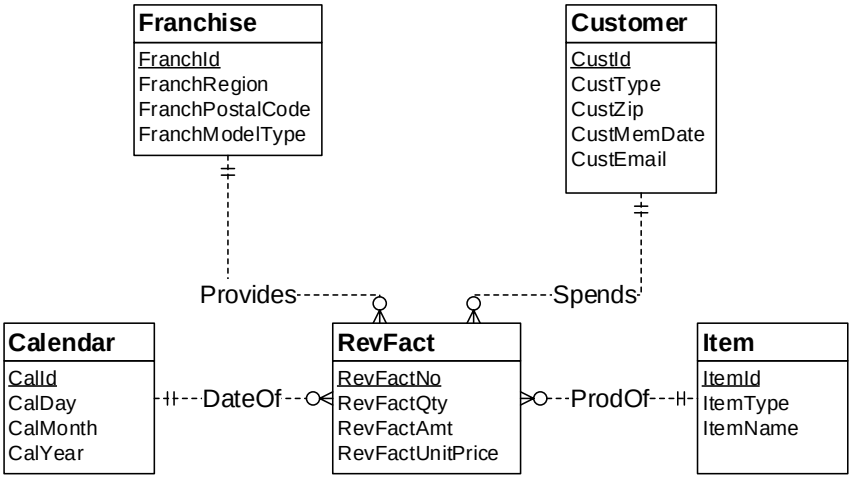


Figure 1: ERD for Fitness Retail Data Warehouse

```
CREATE TABLE Customer (
CustId            INT               NOT NULL,
CustType          VARCHAR2(10)      NOT NULL,
CustZip           INT               NOT NULL,
CustEmail         VARCHAR2(50),
CustMemDate       DATE,
CONSTRAINT CustomerPK PRIMARY KEY(CustId) );

-- Could also add other location columns such as city and state

CREATE TABLE Franchise (
FrachId           INT               NOT NULL,
FranchRegion      VARCHAR2(20)      NOT NULL,
FranchPostalCode  VARCHAR2(10)      NOT NULL,
FranchModelType   VARCHAR2(10)      NOT NULL,
CONSTRAINT FranchisePK PRIMARY KEY(FranchId) );

CREATE TABLE Item (
ItemId            INT               NOT NULL,
ItemType          VARCHAR2(6)       NOT NULL,
```

```
ItemName          VARCHAR2(50)      NOT NULL,
CONSTRAINT ItemPK PRIMARY KEY(ItemId) );

CREATE TABLE Calendar (
CalId       INT   NOT NULL,
CalDay      INT   NOT NULL,
CalMonth    INT   NOT NULL,
CalYear     INT   NOT NULL,
CONSTRAINT CalendarPK PRIMARY KEY(CalId) );

CREATE TABLE RevFact (
RevFactNo         INT               NOT NULL,
FranchId          INT               NOT NULL,
ItemId            INT               NOT NULL,
CustId            INT               NOT NULL,
CalId             INT               NOT NULL,
RevFactQty        INT               NOT NULL,
RevFactAmt        DECIMAL(10,2)     NOT NULL,
RevFactUnitPrice  DECIMAL(10,2)     NOT NULL,
CONSTRAINT RevFactPK PRIMARY KEY(RevFactNo) ),
CONSTRAINT RevFactUnique UNIQUE(ItemId, CustId, CalId) );
```

5. Here are summarizability problems.
   - Incomplete dimension-fact relationship for franchise: The spreadsheet lacks franchise information so additional data collection is required. This data collection should not be difficult as each franchise using a spreadsheet can provide these details.
   - The ERD and spreadsheet indicate that member type is incomplete for members. Some members do not have member types (guests and corporate event customers). This problem can be resolved by default values for guests and corporate customers.
   - Incomplete rollup for location dimension elements because zip codes in member table do not have city and state. More data collection will be necessary to resolve this incompleteness.
   - The membership date applies only to members, not corporate customers and guests. There is no resolution for this incompleteness.
   - No fact-dimension incompleteness for items: events, services, and merchandise have been combined into items so that each revenue fact is associated with an item.

6. The data warehouse tables have been derived from the sample date in the source tables and spreadsheet. The delivery date for the supply purchases uses the default value of the purchase date since the values are missing the source data. New primary key values have been generated for data from the spreadsheet data source.

| ***Calendar*** | | | |
|---|---|---|---|
| *CalId* | *CalDay* | *CalMonth* | *CalYear* |
| 1111 | 10 | 2 | 2021 |
| 1112 | 11 | 2 | 2021 |
| 1113 | 12 | 2 | 2021 |
| 1114 | 13 | 2 | 2021 |
| 1115 | 14 | 2 | 2021 |
| 1116 | 15 | 2 | 2021 |
| 1117 | 16 | 2 | 2021 |

| | | | |
|---|---|---|---|
| 1118 | 17 | 2 | 2021 |
| 1119 | 18 | 2 | 2021 |
| 1120 | 19 | 2 | 2021 |
| 1121 | 20 | 2 | 2021 |
| 1122 | 21 | 2 | 2021 |

| ***Item*** | | |
|---|---|---|
| *ItemId* | *ItemName* | *ItemType* |
| 1111 | Wilson balls | MRCH |
| 1112 | Wilson racket | MRCH |
| 1113 | Adidas shoes | MRCH |
| 1114 | Racket stringing | MRCH |
| 1115 | Ball machine | PASS |
| 1116 | Private lesson | PASS |
| 1117 | Adult class | PASS |
| 1118 | Child class | PASS |
| 1119 | Adult social | EVNT |
| 1120 | Pioneer social | EVNT |
| 1121 | Team practice | EVNT |
| 1122 | Platinum membership | MMBR |
| 1123 | Gold membership | MMBR |
| 1124 | Value membership | MMBR |

- Item identifiers were added for events.
- Item identifiers were changed to integers.

| **Franchise** | | | |
|---|---|---|---|
| *FranchId* | *FranchRegion* | *FranchPostalCode* | *FranchModelType* |
| 1111 | Northwest | 98011 | Full |
| 1112 | Mountain | 80111 | Medium |
| 1113 | Central | 45236 | Limited |

- Franchise identifiers were changed to integers.

| **Customer** | | | | | |
|---|---|---|---|---|---|
| *CustId* | *CustName* | *CustZip* | *CustType* | *CustMemDate* | *CustEmail* |
| 1111 | Joe | 80111 | M1 | 1-Feb-2021 | joe@serv1.com |
| 2222 | Mary | 80113 | M2 | 1-Jan-2021 | mary@serv2.com |
| 3333 | Sue | 80114 | M3 | 3-Mar-2021 | sue@serv3.com |
| 4444 | George | 80112 | M4 | | george@serv4.com |
| 5555 | Frist Data | 80111 | M5 | | |
| 6666 | DU Tennis | 80117 | M5 | | |
| 7777 | Creek | 80111 | M5 | | |

- New customer types were created for non members (M4) and corporate customers (M5)
- Zip codes were added for corporate customers. New data collection is necessary.
- Customer identifiers were added for corporate customers.
- CustMemDate is null (inapplicable) for corporate customers and guests.
- Email addresses were not collected for corporate customers but it may be possible to collect values if desired.

| ***RevFact*** | | | | | | | |
|---|---|---|---|---|---|---|---|
| *RevFactNo* | *CustId* | *CalId* | *ItemId* | *FranchId* | *RevFactQty* | *RevFactAmt* | *RevFactUnitPrice* |
| 1 | 1111 | 1111 | 1111 | 1111 | 2 | $30 | $15 |
| 2 | 1111 | 1111 | 1112 | 1111 | 1 | $200 | $200 |
| 3 | 2222 | 1113 | 1114 | 1112 | 1 | $40 | $40 |
| 4 | 3333 | 1113 | 1113 | 1113 | 1 | $100 | $100 |
| 5 | 4444 | 1114 | 1114 | 1113 | 1 | $40 | $40 |
| 6 | 1111 | 1114 | 1114 | 1111 | 1 | $15 | $15 |
| 7 | 2222 | 1115 | 1116 | 1112 | 1 | $75 | $75 |
| 8 | 4444 | 1116 | 1117 | 1113 | 1 | $150 | $150 |
| 9 | 5555 | 1114 | 1119 | 1111 | 1 | $1,000 | $1,000 |
| 10 | 6666 | 1115 | 1120 | 1111 | 1 | $500 | $500 |
| 11 | 7777 | 1122 | 1121 | 1112 | 1 | $200 | $200 |

- The RevFact table uses data from the Sale, Contains, and ServPurchase tables as well as the event spreadsheet.