Milestone 02

I Overview

A. Motivation

The purpose for this milestone is to create a MDM for our mythical RRV organization. This is based upon the details contained in the "Project Background and Overview" (PBO) document which is available via the Blackboard class-wide project document area and three OLTP data models which will be posted under the per-team File Exchange area **AFTER** sufficient progress has been achieved by the team based on the Project Background and Overview Document. Because we don't have a real organization, and this is merely a class project, the document and OLTP data models will serve as the necessary starting point for creating the MDM. Think of them as the results of some business interviews / facilitated sessions and some of our data audit interviews within our mythical organization. Although you do not NEED to use the QMRT method discussed in class to complete this milestone, the QMRT method is **strongly recommended**. Similarly, part of the rationale for receiving the OLTP models after at least one-week's progress has been achieved is to ensure that the appropriate emphasis is placed on the PBO document (the same reason that OLTP considerations were postponed until step-8 in the QMRT).

B. Scope

Each team will create a single, **STRICT**, star schema MDM that will focus on a **single** business process. This process is **"Sale of a single RRV (instance) to a single Customer (person)"**. We will **ignore inventory measures** for this MDM even though the Background document and OLTP models might contain details about them. Similarly, any other details that don't pertain to the sale of a single RRV to a single Customer will be **ignored**.

C. Asking for Feedback or Help

Since there is no real organization, the instructor will serve as a substitute for the organization with regard to any requirements, questions, or clarifications. If you have general concerns, please feel free to ask about them, but try to get things as far along as you can and ask as specific questions as possible. Please don't hesitate to ask for feedback, but **try to ask direct questions instead of merely submitting a model or set of files and asking "Is this ok?"**
In other words, if you have a specific area of concern, and have done some work, post the file(s) ask directed questions about them! If you have general questions, ask them (but you probably don't need to include any files unless I ask for them as a follow-up).

D. Do's and Don'ts

- **DON'T** INCLUDE ANY SPACES IN THE FILE OR DIRECTORY NAMES FOR YOUR SUBMISSIONS / DELIVERABLES! USE UNDERSCORES!

- **DO** pay attention to the Data Types you choose in the Star Schema, because we plan to use **ROLAP** on **Microsoft SQL Server 2012** as the Physical OLAP Data Storage flavor.

- **DO** integrate Data Types (almost any reasonable data type choice is ok as long as there are **no data-type domain or range issues**.) In general, you will probably use Data Types similar to those used in one or more of the OLTP models, but watch the sizes and lengths.

- **DON'T** be concerned with currency conversions, multi-language support, and internationalization or locale issues.

- **DON'T** be concerned with optimizations -- the focus here is on the domain / range considerations not optimizations.

- **DON'T** attempt to model aggregations. This is beyond scope for this milestone and also a bit premature – and we will be using the tool to generate this for us.

- **DON'T** use Snowflake or other derivative forms. **USE A STRICT STAR SCHEMA**, and remember this is an MDM for a Data Mart **NOT** a Data Warehouse!

- **DON'T** attempt to use Kimball's approach. While his overall philosophy and general advice / MDM patterns and guidelines hold true for us in this project, the Matrix approach would require an enterprise-wide analysis step that is beyond our project scope. In other words, if you tried his approach, it would not give you much advantage since we only essentially have information for a small set of business processes and data marts. And therefore our starting information is not sufficient to bootstrap his methodology in this milestone.

- **DO** remember this is a **STAR SCHEMA** – Denormalization is **crucial**.

- **DO** use "**subject oriented**" **NAMES AND VALUES** for your columns and tables in the MDM.

    - **DON'T** use abbreviations, jargon, or acronyms.

    - **DON'T** use the identical names from the OLTP models.

    - **DO** use longer, more descriptive names (and values) for your columns and tables in the MDM

- **DO** pick or create a naming convention and follow it **consistently** (for example, using either MixedCamelCaseNames or Names_Containing_Underscores for your columns and tables in the MDM). It is also a good practice to use a Prefix on the column names to make it clear which table (and therefore which dimension) they belong. For example, CUS_Country or DLR_Country.

- **DO** use the criteria discussed in the MDM materials as a guide when deciding how to group your dimensional attributes into dimensions and hierarchies rather than thinking that having "as few" dimensions as possible or having "as many" dimensions as possible is somehow desirable.

- **DO** recognize that our project is going to be a **flavor** of Data Warehousing, and as such, some simplifying assumptions will be made. In particular, later milestones will not have very sophisticated Data Staging or ETL systems. Therefore, carrying the Production Keys into the MDM (**BUT NOT AS PRIMARY KEYS**) will be allowed and **encouraged**.

- **DON'T** worry about any DBMS physical storage and optimization considerations (Indexes, Fill Factors, Partitioning of Physical Data Files, File Groups, etc.). These **can all be ignored** (the defaults are fine).

- **DO** worry about the semantic considerations (within reason).

- **DO** consider additional details that are easily calculable based on existing data when exploring the OLTP systems. For example splitting a date or time data type into its constituent parts (a separate year-column, separate month-column, etc.) and modeling the parts as textual columns can greatly simplify things for our purposes. Populating these types of columns is easily accomplished in the later ETL phases via the T-SQL functions available to us. Also, including easily determined details about the date/time is useful (for example, day of year, day of week, weekday, etc. and "textual formats".)

E. Some Final Hints / Rules of Thumb

As a **general guideline** for the **expected size** of your MDM schema, you should be looking at somewhere between **6 to 10 dimensional tables** and roughly **125** (possibly more) **dimensional table columns** (total across all dimensional tables but not including the fact table). I mention this only to provide a "ballpark" range and hopefully prevent teams from under or over thinking the complexity.

DON'T DRAW THE ERD UNTIL AFTER THE IMPORTANT DETAILS (E.G. ADDITIVITY, SCD TYPE, AND IDDS) ARE RELATIVELY STABLE AND FINALIZED. BEFORE THAT, USE THE SPREADSHEETS! IF YOU THINK YOU ARE READY TO CREATE THE ERD, SETUP AN APPOINTMENT WITH THE INSTRUCTOR AND HAVE YOUR M2 ROUGH-DRAFT MEETING!

II Deliverables for your MDM

A. Deliverable-1: (Project Status)

1. Use the project tracking tool / spreadsheet you documented in Milestone 1 to capture the necessary project tasks, estimates, effort, and status for your team.

2. Create a separate PDF report for the current project status at the end of each week. Save the report for each week separately in a single PDF file named "**Txx_M2_D1_Project_Status_Week_Ending_YYYY_MM_DD.PDF**". Notice there are NO SPACES in the filename (use underscores).

   Substitute your two-digit team number for the xx, the four digit year for **yyyy**, the two digit month for **mm**, and two digit day for **dd**.

3. Include all project support files (spreadsheets, etc. using a sensible naming convention) in a subdirectory named "**Txx_M2_D1_Project_Status_files**" within your deliverables zip for this milestone.

Suggestions:

- Separate spreadsheet files would make this more modular than trying to edit a single spreadsheet as a team.

- In the PDF file, be aware of the formatting, and you might also want to play with the "Page Setup" to see if Portrait or Landscape works better for you.

- If information needs to span multiple pages, use page breaks and repeated column headings / row headings to make the presentation of the information usable.

- You **DON'T** need to include any "backup files" or previous versions (just the final version of the deliverable files).

B. Deliverable-2: (DIM and FACT Metadata of the MDM)

1. Use the data modeling / spreadsheet tool you documented in Milestone 1 to capture the necessary metadata for your model. At a **MINIMUM**, create one separate sheet (tab or file) for the Fact table and one separate sheet (tab or file) for each Dimension table (e.g. if you had 5 dimensions, there would be 1+5 = 6 sheets / tabs).

2. For the Fact table, include the name of the table and important metadata for each FACT table column (this includes all primary key and foreign key columns). Important metadata should include (**at a minimum**) column name, data type, nullability, description, **and additivity**.

3. For each Dimension table, include the name of the table and important metadata for each Dimension table column (this includes all primary key and foreign key columns). Important metadata should include (**at a minimum**) column name, data type, nullability, sample value, and description.

4. Create a PDF report for this data model, and save it in a single file named "**Txx_M2_D2_MDM_metadata.PDF**" (substitute your team number for the Txx). This report must contain all the information for the fact table and each dimension table (as mentioned above).

5. Include all source files (spreadsheets, etc. using a sensible naming convention) in a subdirectory named "**Txx_M2_D2_MDM_metadata_files**" within your deliverables zip for this milestone.

Suggestions:

- Separate spreadsheet files would make this more modular than trying to edit a single spreadsheet as a team.

- In the PDF file, be aware of the formatting, and you might also want to play with the "Page Setup" to see if Portrait or Landscape works better for you.

- If information needs to span multiple pages, use page breaks and repeated column headings / row headings to make the presentation of the information usable.

- You **DON'T** need to include any "backup files" or previous versions (just the final version of the deliverable files).

C. Deliverable-3: (IDD Details for all DIMs)

1. Because an ERD cannot capture a MDM completely, you must include an Intentional Dimension Diagram (IDD) for each and every dimension in your MDM (and all their hierarchies). There is no "standard format" for this and there is also no "specific tool" for creating this -- having said that, I recommend using Excel or a similar spreadsheet tool. Look at the examples below as well as the BB, articles, and lecture slides for suggestions you might want to incorporate in your "diagrams". Make sure that the level information clearly captured. Always list the finest level of granularity as the bottom level (last on the page / sheet / diagram).

2. Create a PDF version of each IDD using the tool's native facilities or by printing to a PDF file via PDF Creator or similar technique. As always, ensure that the PRINTED FORMAT used in the PDF is legible and useful.

3. Submit all files required / generated by the tool and the PDF files as described below.

Suggestions:

- Be sure to clearly indicate the "Main" SCD Type for the DIM or equivalently the SCD Type for each and every DATT / PROP in the DIM.

- Indicate the **HIER Names**, **LVL Names**, **MBR Keys**, **MBR Names**, Assumptions, Estimated Member cardinality numbers / formulas and **MBR Property details** as well as an obvious indication of the hierarchical (tree-like) relationships between the LVLs for all the DIMs, HIERs, and LVLs you define.

- **DON'T** show **EXTENSIONAL** VIEW. SHOW THE **INTENTIONAL** VIEW. This is an IDD not an EDD!

- Regardless of the technique you use to represent this, create a separate PDF file for each IDD. Make sure each of these files contains all the information. Please feel free to use Word, Excel, or any "useful tool" to create and edit the IDD more easily.

WHAT TO DO for M2-Deliverable-3:

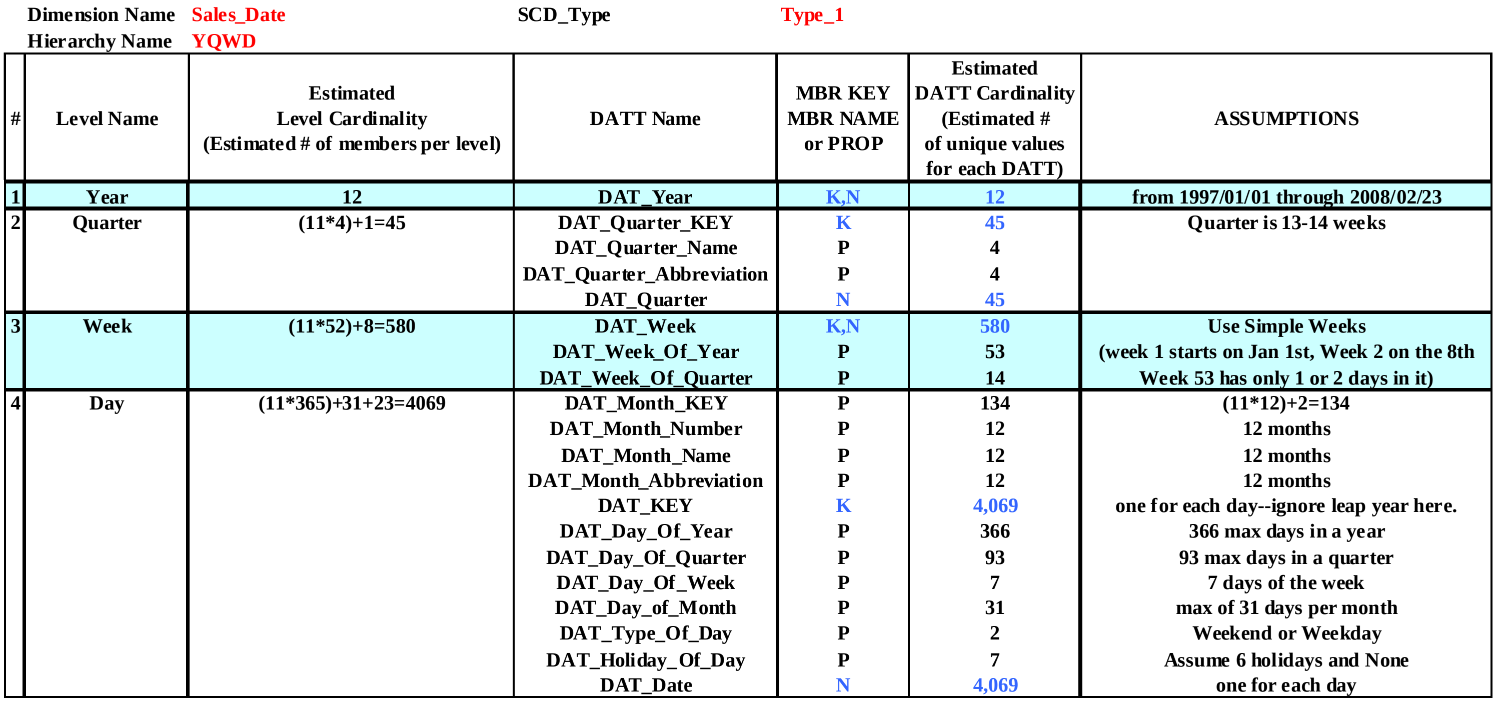
Generate a single PDF file for each IDD, containing the printable / viewable format of the IDDs for each DIM.

Name each PDF file "Txx_M2_D3_IDD_<dim-name>_<hier-name>.pdf", substituting the appropriate team number, dimension name, and hierarchy name.

Also include whatever source files (*.xls, *.doc, etc.) used to create the IDDs in a directory named "Txx_M2_D3_IDD_files".

For creating the IDD, I recommend using the format provided as an example in lecture. Look at the files posted to the blackboard for the MDM lectures. Any reasonable format that includes all the necessary details is fine, but if you are using something very different from the examples given in class, Please check with the instructor before you put too much effort into creating the IDDs.

For example, if we had a Sales_Date dimension with a YQWD Hierarchy whose levels were based on Year, Quarter, Week, and Day we could show this information like this:



Dimension Name Sales_Date SCD_Type Type_1
Hierarchy Name YQWD

#	Level Name	Estimated Level Cardinality (Estimated # of members per level)	DATT Name	MBR KEY MBR NAME or PROP	Estimated DATT Cardinality (Estimated # of unique values for each DATT)	ASSUMPTIONS
1	Year	12	DAT_Year	K,N	12	from 1997/01/01 through 2008/02/23
2	Quarter	(11*4)+1=45	DAT_Quarter_KEY	K	45	Quarter is 13-14 weeks
			DAT_Quarter_Name	P	4	
			DAT_Quarter_Abbreviation	P	4	
			DAT_Quarter	N	45	
3	Week	(11*52)+8=580	DAT_Week	K,N	580	Use Simple Weeks
			DAT_Week_Of_Year	P	53	(week 1 starts on Jan 1st, Week 2 on the 8th
			DAT_Week_Of_Quarter	P	14	Week 53 has only 1 or 2 days in it)
4	Day	(11*365)+31+23=4069	DAT_Month_KEY	P	134	(11*12)+2=134
			DAT_Month_Number	P	12	12 months
			DAT_Month_Name	P	12	12 months
			DAT_Month_Abbreviation	P	12	12 months
			DAT_KEY	K	4,069	one for each day--ignore leap year here.
			DAT_Day_Of_Year	P	366	366 max days in a year
			DAT_Day_Of_Quarter	P	93	93 max days in a quarter
			DAT_Day_Of_Week	P	7	7 days of the week
			DAT_Day_of_Month	P	31	max of 31 days per month
			DAT_Type_Of_Day	P	2	Weekend or Weekday
			DAT_Holiday_Of_Day	P	7	Assume 6 holidays and None
			DAT_Date	N	4,069	one for each day

WE ARE MUCH MORE CONCERNED ABOUT CONTENT THAN STYLE!
When in doubt, follow the conventions and choose function over style.

D. Deliverable-4: (ERD of the Physical Data Model for the MDM)

1. Use the drawing / data modeling tool you documented in Milestone 1 to create your strict, star schema ROLAP MDM **physical data model** using MS SQL 2012 as the physical database for this deliverable

2. Create a PDF version of the PDM / ERD for this data model. The diagram must include the fact table, and all dimension tables, all references (using crows-feet notation). The diagram must also include for each table, all column-names, and an indicator of all primary key / foreign key columns.

3. .Create a PDF file for this ERD, and save it in a single file named "**Txx_M2_D4_PDM_erd.PDF**" (substitute your team number for the Txx).

4. Include all source files (the raw diagramming files, etc. using a sensible naming convention) in a subdirectory named "**Txx_M2_D4_PDM_erd_files**" within your deliverables zip for this milestone.

Suggestions:

- You **DON'T** need to include any "backup files" or previous versions (just the final version of the deliverable files).

- **DON'T** try to fit each diagram all on a single page, I recommend using two pages for the ERD.

- **DON'T** forget to ensure that all the names in all the tables and columns can be seen!

- **DO** look at the page formatting. You might also want to play with the "Page Setup" to see if Portrait or Landscape works better for you.

- **DO** put the Fact table in the "middle" of the diagram, but **DON'T** have it spanning the page break (In other words, put the fact table on one page or the other, not both).

III Rough Draft Meeting Request and Files Submission

Each team should submit a rough draft meeting request (see the blackboard for the separate assignment deliverable). This is a request for the team to meet with me to review the current state of your model / metadata (i.e. the FACTs, DATTs, DIMs, IDDs, etc.). It is intended to contain a reasonably substantial set of content, but also intended to ask / answer questions about the model (as it exists at this point).

The meeting request has a single, simple deliverable (filling out the form for the meeting request).

The Rough Draft Meeting Files submission is a separate submission, to be done before (or after) the actual meeting. This submission should follow the same format / layout conventions as the final submission but it will not have all the details or files. As a minimum it should include the current project tracking data, and any relevant details / questions captured. In other words, this is NOT an ERD (it is too soon for that), but it might be a set of spreadsheets containing the current potential DATTs / FACTS, the DIM details, and some initial IDDs.

See the next section for format / layout / naming convention.

IV Final packaging for M2 submissions:

WHAT TO DO for the Final M2 submission:

Create a directory named Team_xx and place all files and subdirectories within it. Create a SINGLE ReadMe.txt file, briefly listing what files and file formats are included in submission as well as any other relevant details if necessary. Create a SINGLE CoverSheet.txt file, briefly listing the Milestone number, Team number, CLASS-IDs of all team members who worked on this milestone, and the submission date. Place the directories and files named in the deliverables in this Team_xx directory and then use 7zip to zip the Team_xx directory into a SINGLE ZIP ARCHIVE FILE named "Team_xx_M2_ROUGH_DRAFT_Deliverable.zip" for the rough draft submission and "Team_xx_M2_FINAL_Deliverable.zip" for the final submission.

Inside the SINGLE Zip file that you submit, it should look something like this:

```
📁Team_01
    |
    +–📁 T01_M2_D1_Project_Status_files
    |
    +–📁 T01_M2_D2_MDM_metadata_files
    |
    +–📁 T01_M2_D3_IDD_files
    |
    +–📁 T01_M2_D4_ERD_PDM_files
    |
    +–📁 T01_M2_Other_Support_Files
    |
    +– T01_M2_D1_Project_Status_Week_Ending_2017_02_26.PDF.PDF
    |        (one for each week of progress)
    |
    +–T01_M2_D2_MDM_metadata.PDF
    |
    +–T01_M2_D3_IDD_Sales_Date_YWQD.PDF
    |        (separate one for each DIM and each hierarchy)
    |
    +–T01_M2_D4_PDM_erd.PDF
    |
    +–CoverSheet.txt
    |
    +–ReadMe.txt
```

The goal for the rough draft is to make significant progress towards the final deliverable and setup an appointment to discuss specific questions with respect to the current progress on the model.

Teams that make incremental progress and attempt regular communication are more likely to be successful in reaching the final deliverable.
Stepwise refinement is always easier—☺.

SUBMIT (NOT SEND!) THE DELIVERABLE ZIP FILES
TO THE APPROPRIATE ASSIGNMENT AREA ON THE BLACKBOARD
BEFORE DATE AND TIME THAT IT IS DUE!

FOLLOW THE FORMAT AND NAMING RULES!
ONE SUBMISSION PER TEAM!
DO NOT PRINT ANYTHING!
DO NOT EMAIL ANYTHING!
DO NOT TURN IN ANY HARDCOPY!