CALIFORNIA STATE UNIVERSITY EAST BAY

Course **ITM 6280 – Data Warehousing**

Term **Spring Quarter 2018**

Homework 1

Total points: 25 Due: May 13, 2018

In this homework you will learn how to create a simple relational database using real life data. You will start with saving the **Movie2010_11_Data.mdf** file in SQL server. You will create two dimensional tables – dimMovie and dimDate. You will then use SQL server integration services (SSIS) to populate these tables using the saved Fact table data in Movie2010_11_Data. Next, you will use SQL server Data Tools (SSDT) to create a project where you will select these three tables and create relationships. In the end you will use this relationship database to create a mining model.

Add Movie2010_11_Data in SQL server

1. Download Movie2010_11_Data.mdf from blackboard and save it in DataBase Files (if you are using BayCloud) or in Data folder (if you have downloaded SQL server and its components) inside MSSQL which contains other mdf files.
2. This data is a scraped from **BoxOfficeMojo.com**. It contains details of top grossing movies on each day from 2010 to 2011. Following is the website.
   http://www.boxofficemojo.com/daily/?view=year&yr=2010&p=.htm
3. Open SQL server Management Studio (SSMS) and connect to the SQL server engine and Analysis Services. Then attach this database to the server.

Creating Dimension Tables

4. Expand the Movie2010_11 database to see the columns of **dbo.Fact** table. Right click Tables to create dimension tables **dimMovie** and **dimDate**.
5. dimMovie should have following attributes **Movie_PK, MovieName, Gross, Budget, MPAA, and Genre. Remember to match the data types and size with the attribute in the Fact table.** For example, Movie_PK should be varchar(255). Uncheck the Allow Nulls box for **Movie_PK.** Select Movie_PK and right-click to set it as a primary key
6. Similarly create a **dimDate** table with columns **Date_PK, varYear, varMonth, varDate**, **varDay, and Total_Days** with appropriate data type. Set Date_PK as the primary key in

dimDate and uncheck Allow Nulls. Expand the **Identity Specification** in the column property of Date_PK attribute. Change **Is Identity** to **Yes**.

Populating Dimension Tables

7. Open Visual Studio and create a new project. In the New Project window select **Integration Services Project** from **Integration Services**. You can name the project as **Populating Tables**.
8. Drag **Data Flow Task** from **SSIS Toolbox** to the **Control Flow** window.
9. Right-click the **Connection Manager** window to create two new database connections for the project. Select Movie2010_11 as the database. Rename one connection as **Source** and other as **Destination**. Note that both source and destination tables are in the same database.
10. Click on the **Data Flow** tab and drag the **OLE DB Source** from **SSIS toolbox** to the Data Flow window. Rename the source as **Movie Source**. Double click it to open the editor and select the **Fact** table as the source.
11. Drag the **Slowly Changing Dimension** (SCD) from SSIS Toolbox and connect it to the Movie Source.
12. Double click SCD to open the wizard and select **dimMovie** as the **Destination** table. This will match the input columns of Fact table to the dimension columns of dimMovie table. Select **ID** as the input column for **Movie_PK** dimension column and set it to **Business Key**. Select all the other columns of the dimMovie table as **Changing attribute**. Finish the wizard.
13. Drag **Multicast** tool from the SSIS toolbox and connect it to the Insert destination tool using **red arrow**. In the **Configure Error Output** window, select the Redirect row from both drop down arrows.
14. Click the **Start** tab with a green arrow on top ribbon. This will start the flow of the data from the Source table (Fact) to the destination table (dimMovie). After successful run, the dimMovie table will be populated with the data. **Take a screenshot of the successful run with green check marks against the tools.**
15. Click on the **blue hyperlink** below the Data Flow window to stop debugging.
16. Use the same model to populate dimDate table. Set **varDate, varDay, varMonth, and varYear** as **Business key**. Combination of these attributes will create a composite alternate primary key for consecutive years. Set **Total_Days** as **Changing attribute**.
17. Open the Management Studio and right-click dimMovie and dimDate to select top 1000 rows. **Take a screenshot of the populated data in dimMovie**.

18. Right-click **Database Diagrams** in Movie2010_11 database and select **New Database Diagram**. Select all three tables. Drag and drop the primary keys over foreign keys to create relationships. Save the diagram as **Relationship**.

Creating Relationships in Analysis Services

19. Open **SQL Server Data Tools** and create a new **Analysis Services Tabular** project named **Homework**.
20. By right clicking Data Sources, add Movie2010_11 as the database source. You can use Homework as a Friendly connection name.
21. Select all three tables. From dimFact only filter **ID, Top10_Gross and Date_ID** columns. Import all these tables with filtered columns.
22. You will find all the three tables populated in the **Model.bim** window. Click on the **Diagram** view in the lower right corner of the Model.bim window.
23. Right click the table header to create relationships. **Take a screenshot of the relationships.**

Data Mining

24. Create a new **Analysis Services Multidimensional Project** in SSDT named **Time Series**.
25. In the solution explorer, right click **Data Sources** and create a new data connection **Movie2010_11**. Data source name should be same.
26. Right click **Data Source View** to add Movie2010_11 Relational Data Source.
27. Include all the tables except sysdiagrams and complete the wizard.
28. Right click **Mining Structure** to open the wizard. Check From existing relational database or data warehouse.
29. Select **Microsoft Time Series** as the mining model.
30. Select **Fact** table as the case. Select **Date_ID** as key and **Top10_Gross** as both input and predictable. Finish the wizard.
31. Start the deployment and processing by clicking the **Start** tab with a green arrow on top ribbon.
32. After successful deployment and processing of the mining model, click on the **Mining Model Viewer in Fact.dmm tab**.
33. **Take a screenshot of the time series chart**.

Questions:

1. **Why Multicast tool is used? What will happen if it is not used?**
2. **How many records are used as an input to the dimMovie table and how many records got populated? Why all records are not populated in the dimMovie table?**
3. **What are the cardinalities of the relationships?**
4. **What do you infer from the time series chart? Why there are periodic spikes in the chart?**

Deliverables:

1. **Submit a screenshot of the successful run of the SSIS model.**
2. **Submit a screenshot of the SSMS showing the at least last 5 records of the dimMovie Table.**
3. **Submit a screenshot of the schema of the Homework database in the Diagram view showing the relationships.**
4. **Submit the screenshot of the time series chart.**
5. **Upload this screenshots and answer the questions in Blackboard.**