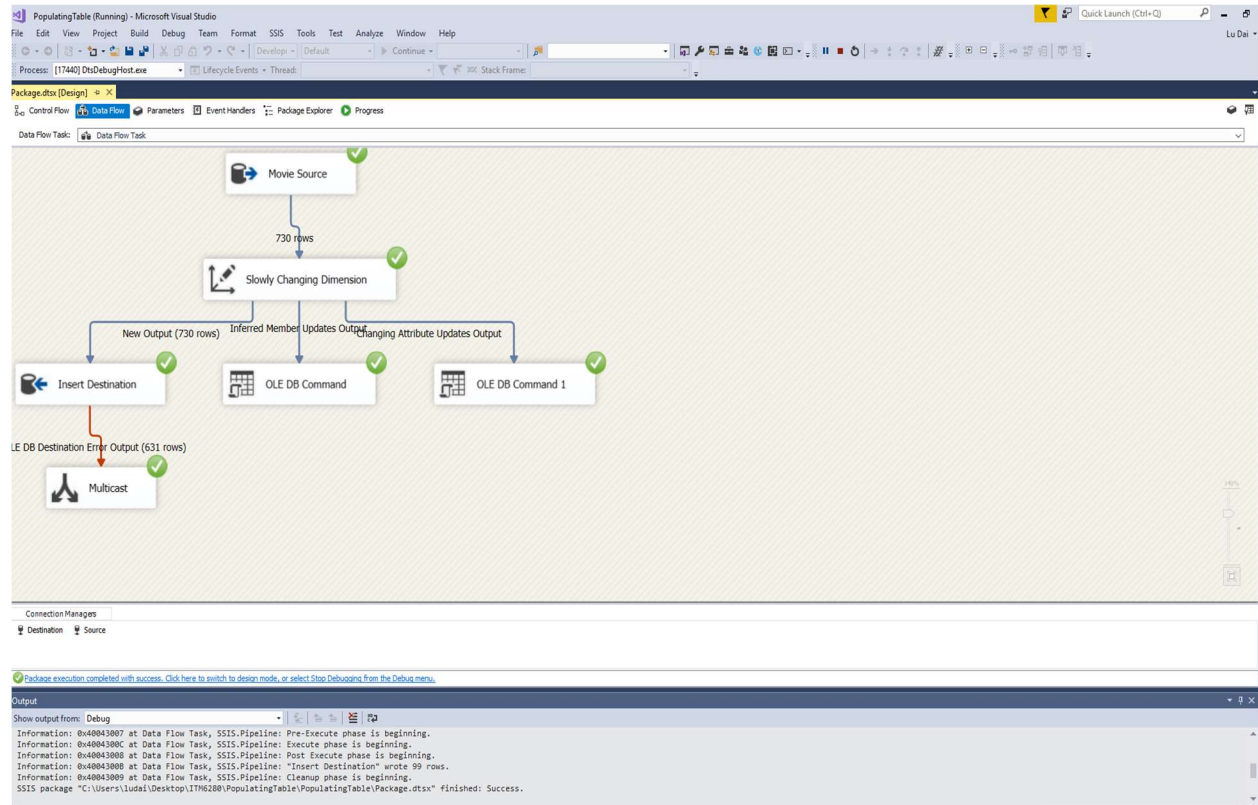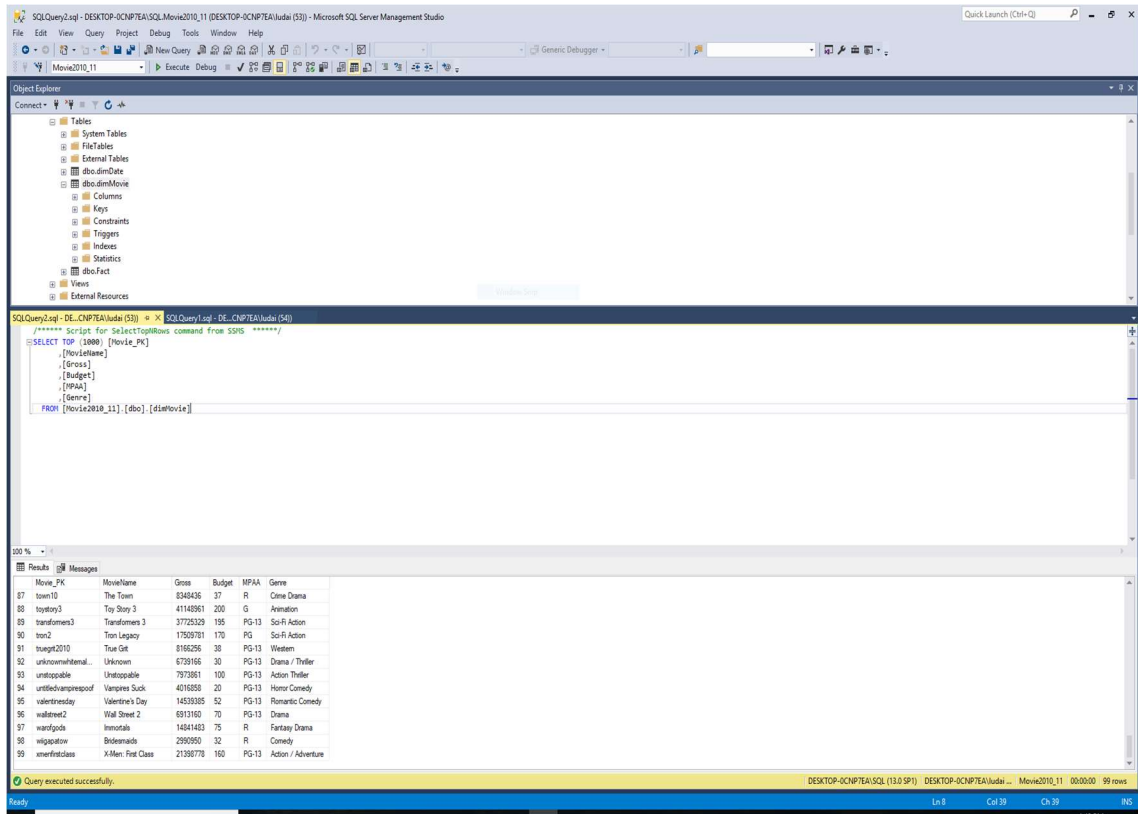**Name: Lu Dai (qx8329)**

Class: ITM6280

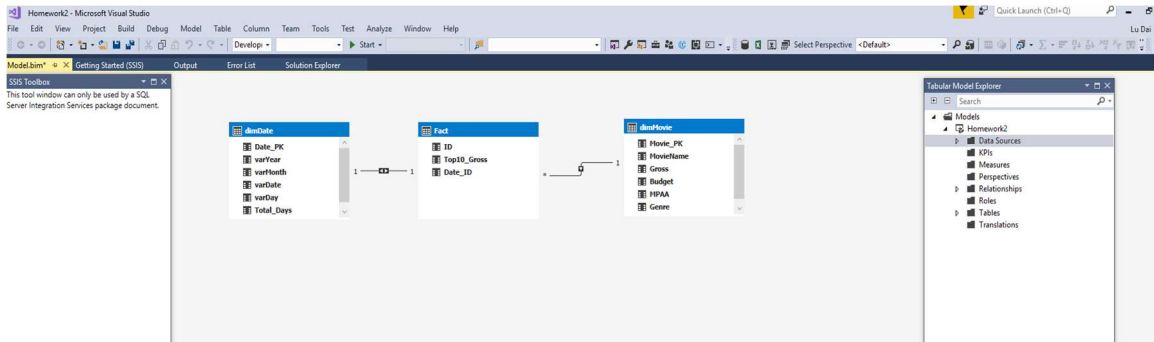**Homework Assignment: Homework 1**

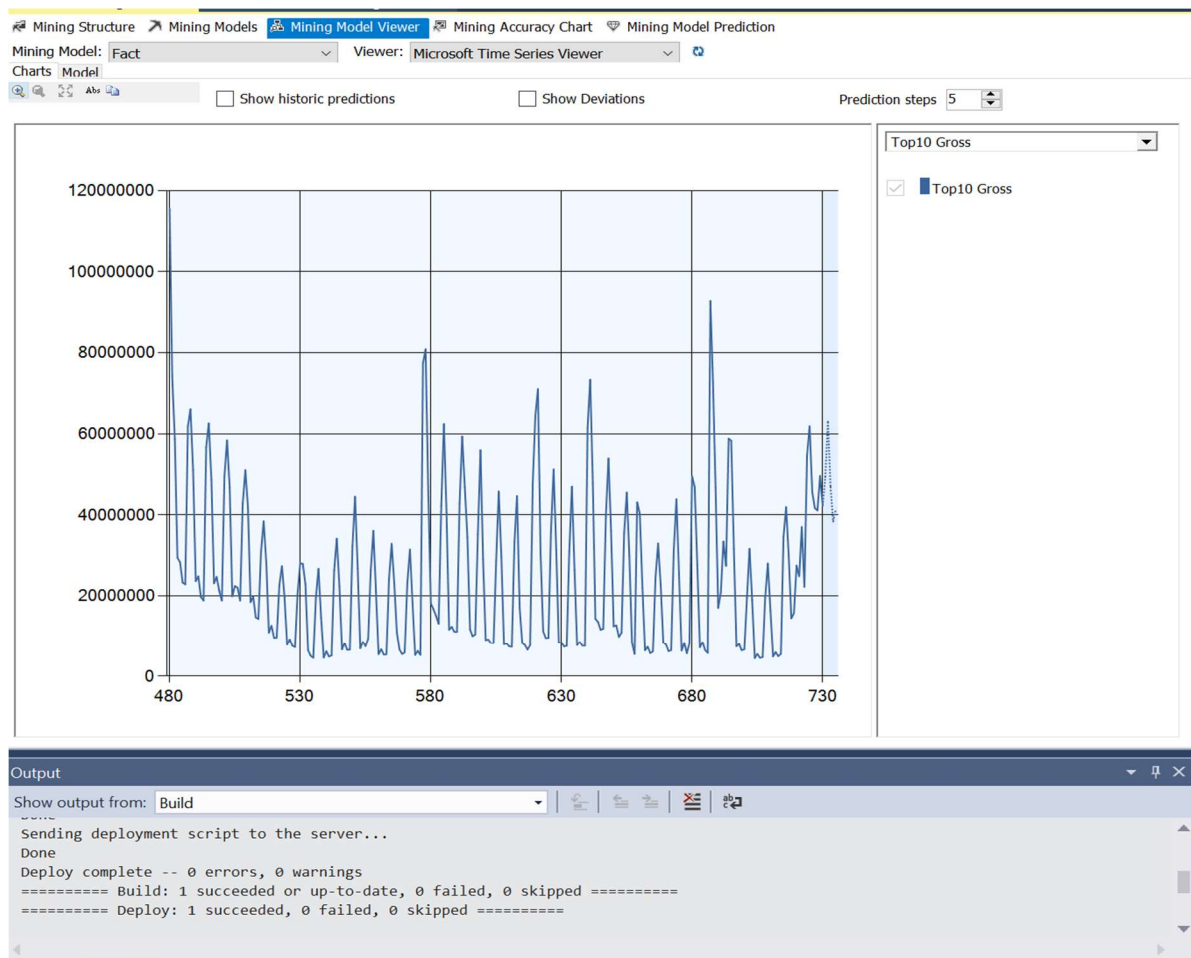1. **Submit a screenshot of the successful run of the SSIS model.**

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.**

Questions:

1. Why Multicast tool is used? What will happen if it is not used?

The reason Multicast tool is used is to create or distribute exact copies of the source dataset to one or more destination datasets. Multicast Tool is very useful when we want to apply different transformations on the same dataset in parallel. Also, Multicase tool can filter the data.

If Multicast is not used, we cannot pass the same data to multiple outputs, spend much more time in populating dimension tables, and may get the data which were filtered out.

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?

In this case, 730 records are used as input to the dimMovie table, and 99 records got populated. Because we used the changing attribute to update 631 rows, only 99 rows were populated.

Another reason is that we handle data flow errors by using error redirection. When a data flow component applies a transformation to column data, extracts data from sources, or loads data into

destinations, errors can occur. During populating data, we select the Redirect row to ignore errors in the configure error output window. That is why populated records are less than inputted records.

3. What are the cardinalities of the relationships?

The cardinalities of relationship is the number of occurrences in one entity which are associated (or linked) to the number of occurrences in another.

There are three cardinalities of relationship, known as:

one-to-one (1:1)

one-to-many (1:M)

many-to-many (M:N)

One-to-one (1:1) is where one occurrence of an entity relates to only one occurrence in another entity.

One-to-Many (1:M) is where one occurrence in an entity relates to many occurrences in another entity.

Many-to-Many (M:N) is where many occurrences in an entity relate to many occurrences in another entity.

In this project, one Date_PK in dimDate table can only relates to one Date_ID in fact table. So, it is one-to-one relationship. One Movie_PK in dimMovie table can relate to many ID in fact table. Therefore, it is one-to-many relationship.

4. What do you infer from the time series chart? Why there are periodic spikes in the chart?

In this time series chart, I find that Top10_Gross indicate seasonality pattern. The median of Top10_Gross in these two years is about $40,000,000. Moreover, I find that the periodic spikes always repeat every seven days. I infer that more people like to go to cinema on weekend, which significantly increases the gross.

Teacher feedback:

Apart from splitting the data to multiple destinations, Multicast also check for errors. For example, if there are duplicate entries then Multicast redirect the record rather than populating it. This is primary reason why we are using Multicast in our model.

Why 631 records were treated as errors and not populated in dimMovie? The main reason why not all movieIDs are populated is because movieID have duplicate entries in the Fact table. Several movies remain the top grossing movie for several days. So, when we populate the movieID in the movie_PK all the duplicate movieIDs get redirected by the Multicast tool and we get less but unique records.