INFO 7290 (DWBI): HW05a – Analyze Data

Smit Patil

1. Look at the page above and provide a brief (less than 1/2 page overview of how you will go about loading this data). Do not copy-paste instead give an idea of what you think the steps will be to load the data and how you will go about it.

The COVID-19 Nursing Home dataset consists of data reported by the nursing homes to the Centers for Disease Control and Prevention (CDC) or, more specifically, the National Healthcare Safety Network (NHSN) system. The dataset contains about 322k rows and 91 columns. It can be downloaded using various file formats, including the popular CSV and TSV file formats. I will be using CSV formatted file to load the data into the SQL Server.

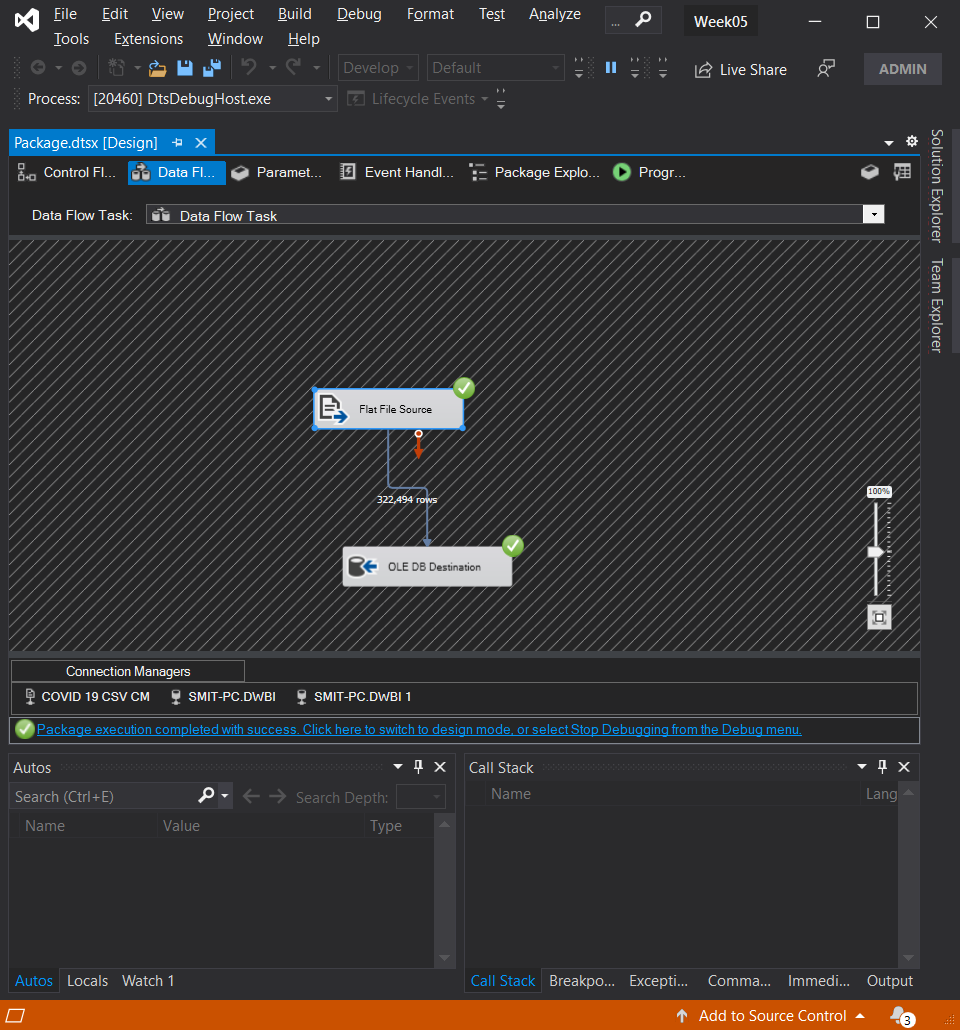
Before loading data in SQL Server, we have to create a SQL Server table to load our data. The table should be similar to that of our file. After creating the table in SQL, we can load the data using multiple ways. The two ways I will be loading the data is by using SSIS and BCP.

To load the data using SSIS first, we must create a data integration package in Visual Studio. We have to complete a data flow task in our control flow. The data flow task will mainly contain two operations. First, to load the flat file data, CSV file in our case. Second, ingest the data into the table created in SQL by creating an OLE DB connection. We will select our CSV files in our flat-file connection and check whether all our columns are loaded correctly. After our file is loaded successfully, we will connect it with our table in the database and review it for the columns' mappings. We might need to edit the mappings as per the requirements. We also need to check for the columns' data type to be loaded without errors.

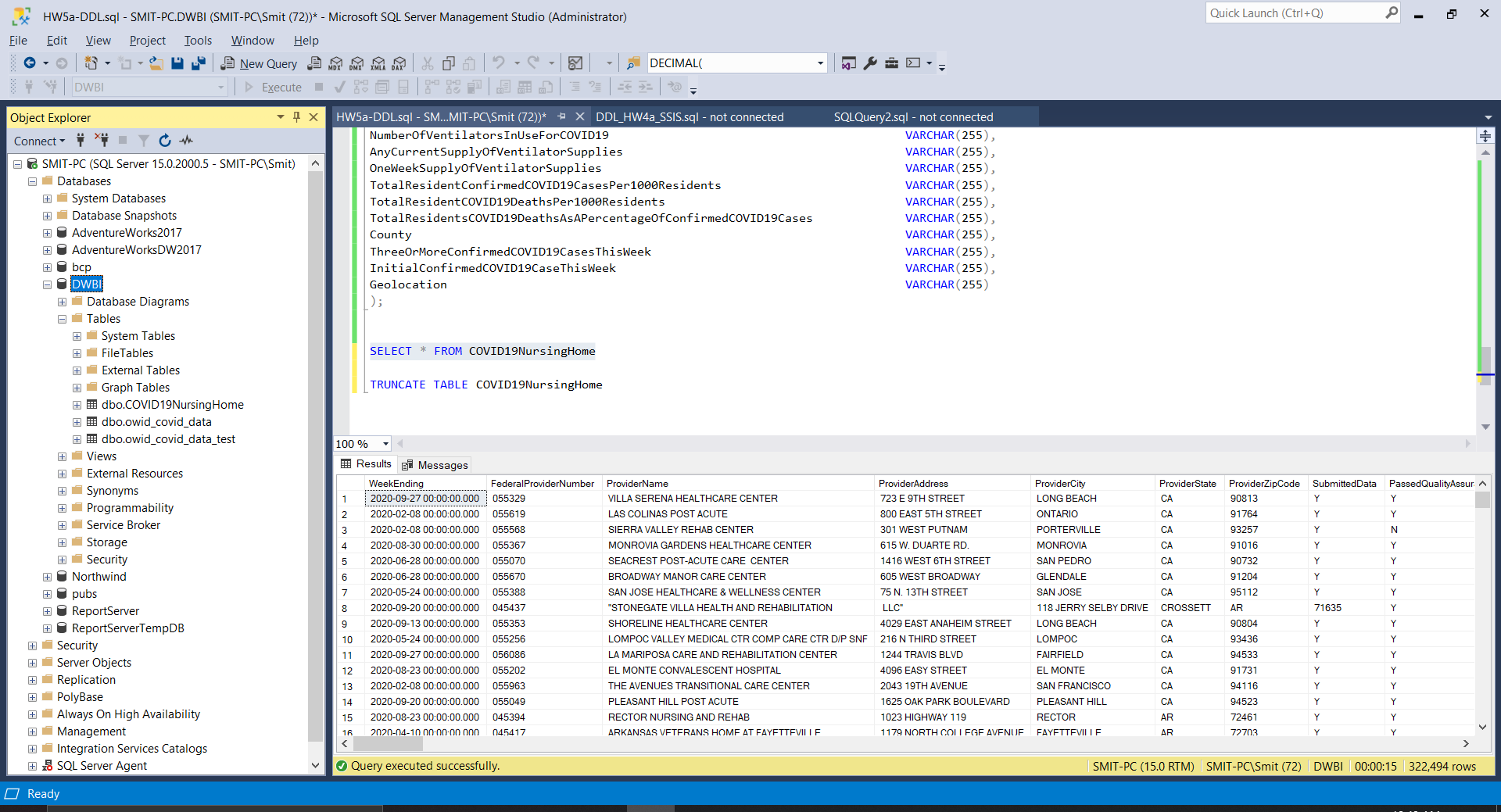
To load the data using BCP, we have to open the command prompt and use the following **"bcp *"destination table"* in *"source file"* -S *"server name"* -d *"database name"* -T"** BCP command to create a format file. After making the format file, we can use that format file to set our data types, field terminators, and row terminators as required. Then we can use the following **"bcp *"destination table"* in *"source file"* -S *"server name"* -d *"database name"* -T -f "format file""** command to load our data.

1. Load the data into SQL Server using SSIS

SSIS package running successfully and loading the data in the SQL Server database:

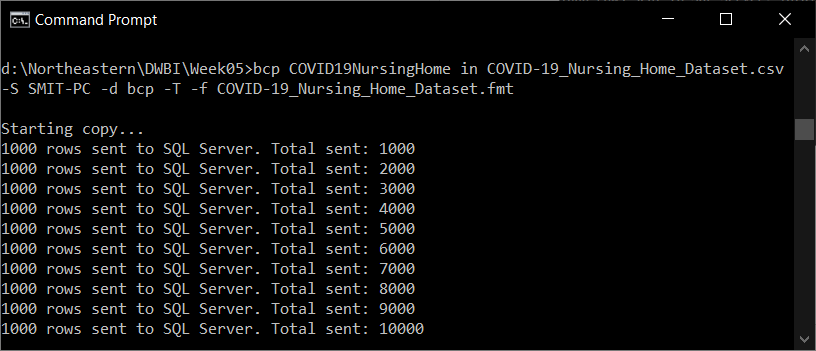


The output of the table in SQL Server showing the data is correctly

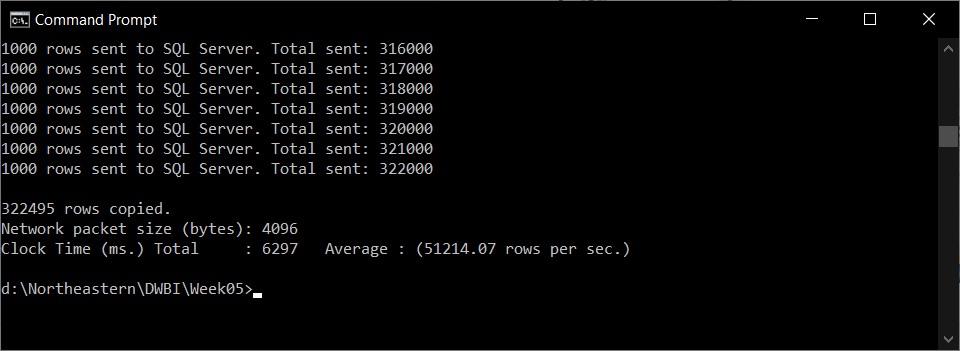


1. Load the data using BCP

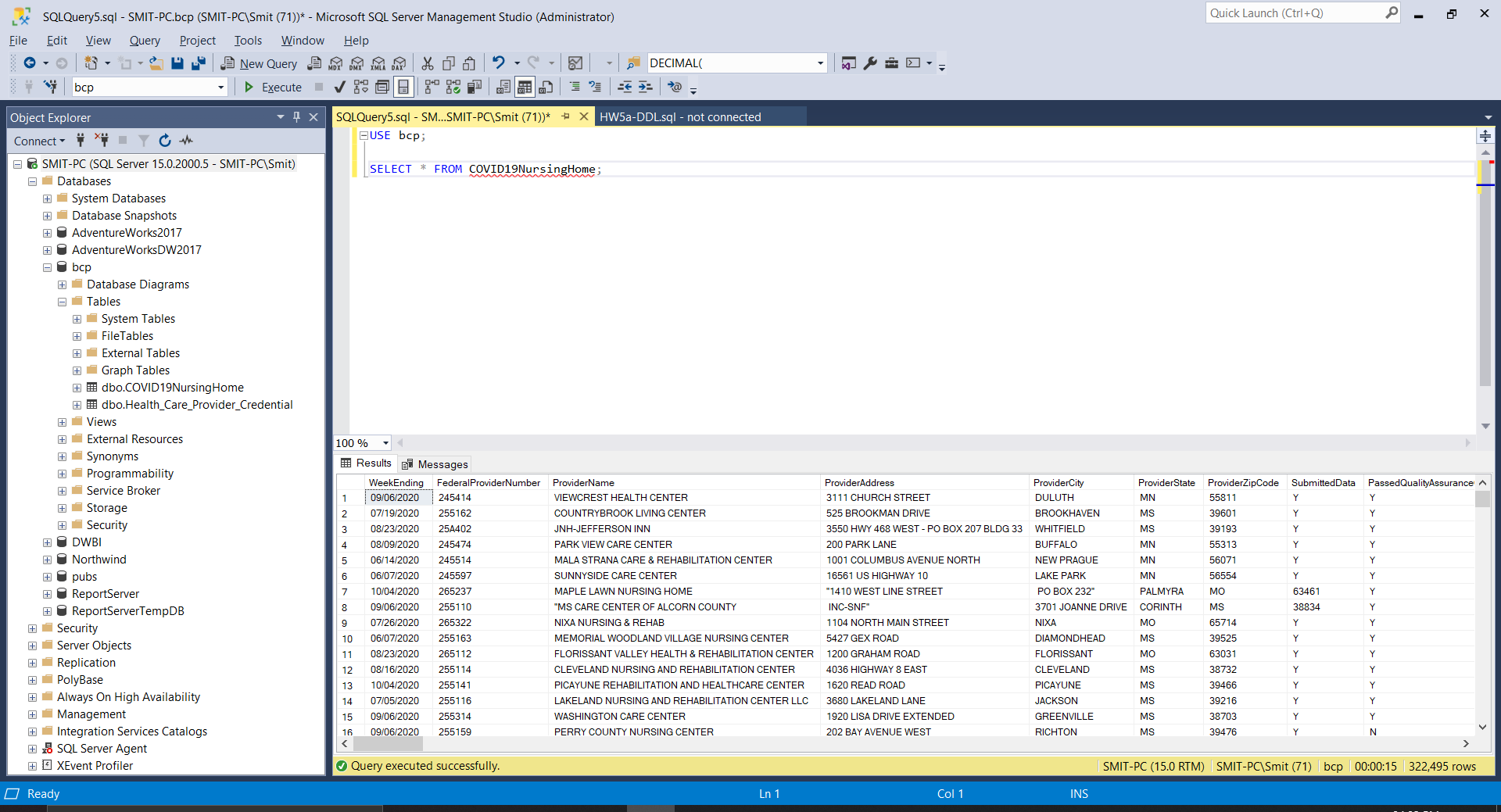
Using the format file to load the data into the database



Contd.



The output of the BCP database in SQL showing that all the rows are loaded successfully



1. Analyze the data - After loading the data provide a set of analyses on the data. Be as complete as possible (about a page or so with a table or two).

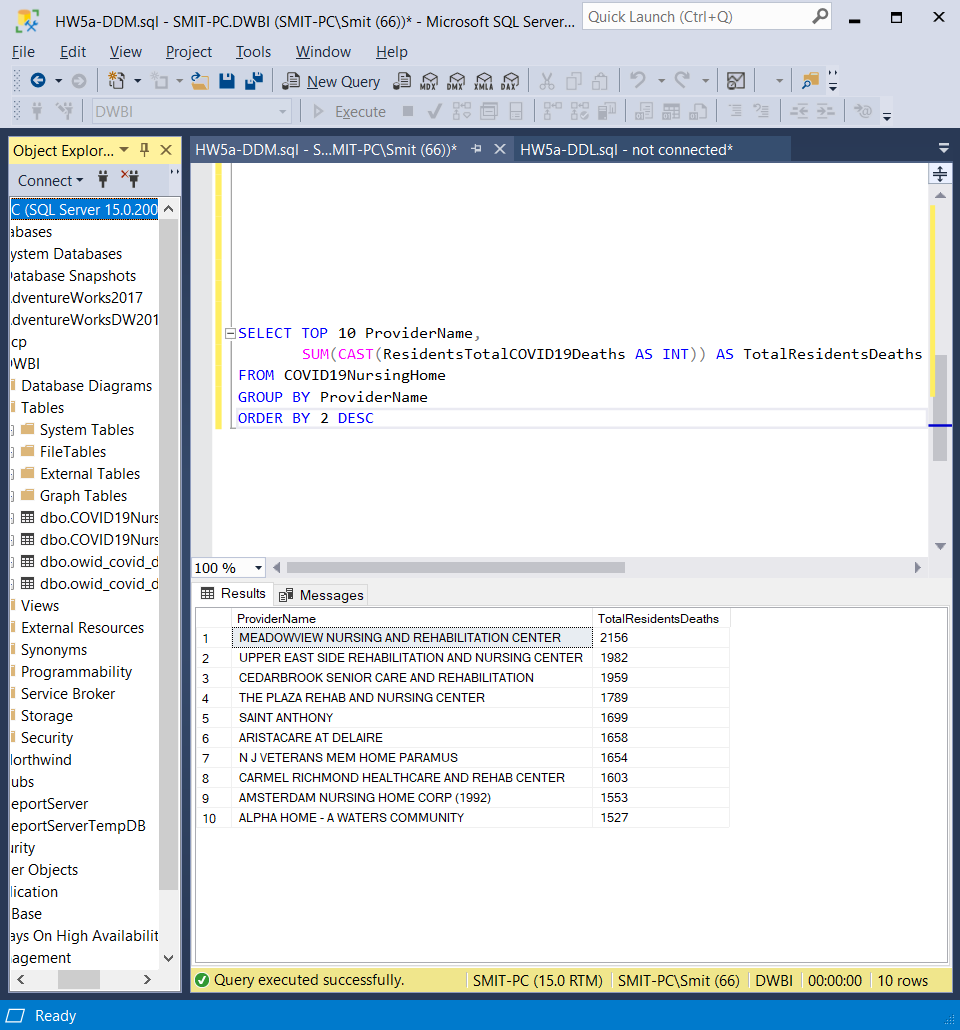
SELECT TOP 10 ProviderName,

SUM(CAST(ResidentsTotalCOVID19Deaths AS INT)) AS TotalResidentsDeaths

FROM COVID19NursingHome

GROUP BY ProviderName

ORDER BY 2 DESC



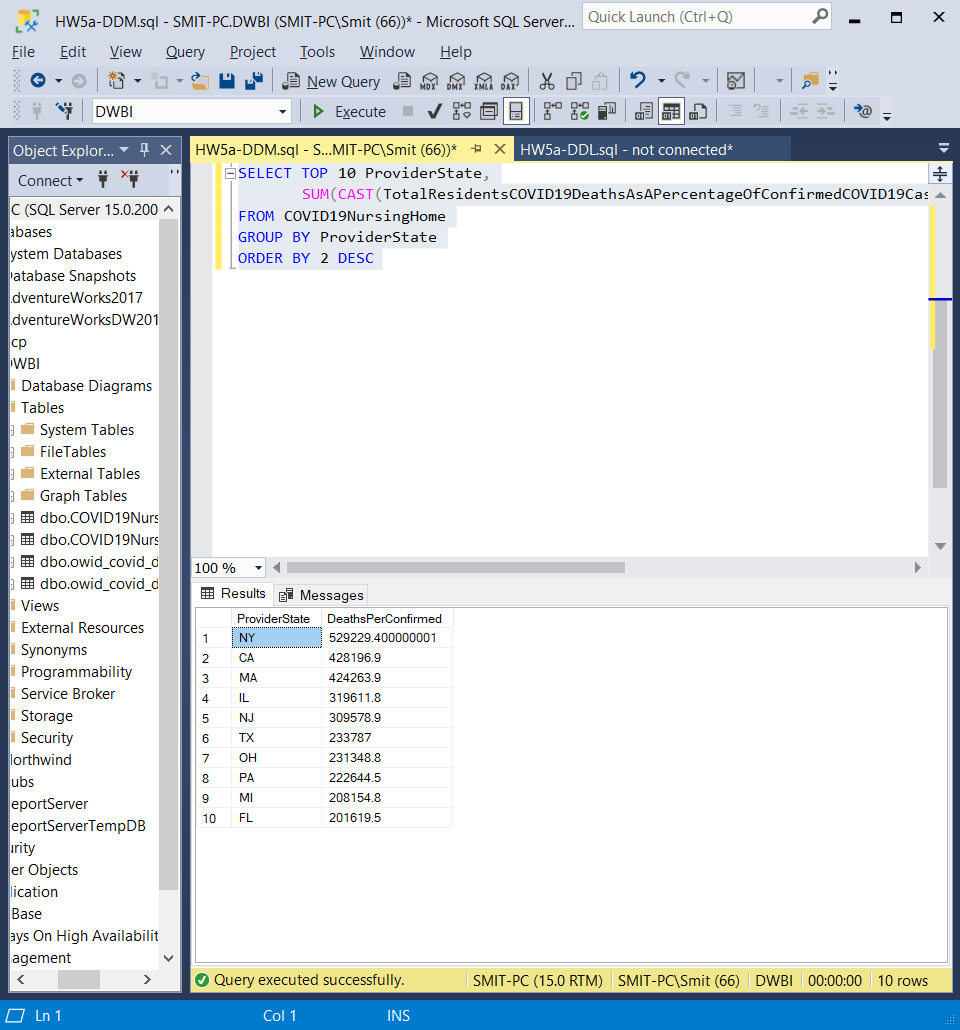
SELECT TOP 10 ProviderState,

SUM(CAST(TotalResidentsCOVID19DeathsAsAPercentageOfConfirmedCOVID19Cases AS FLOAT)) AS DeathsPerConfirmed

FROM COVID19NursingHome

GROUP BY ProviderState

ORDER BY 2 DESC



Map

Description automatically generated

Chart, histogram

Description automatically generated

1. Assume you would get this file again next month what steps would you take to load the data again

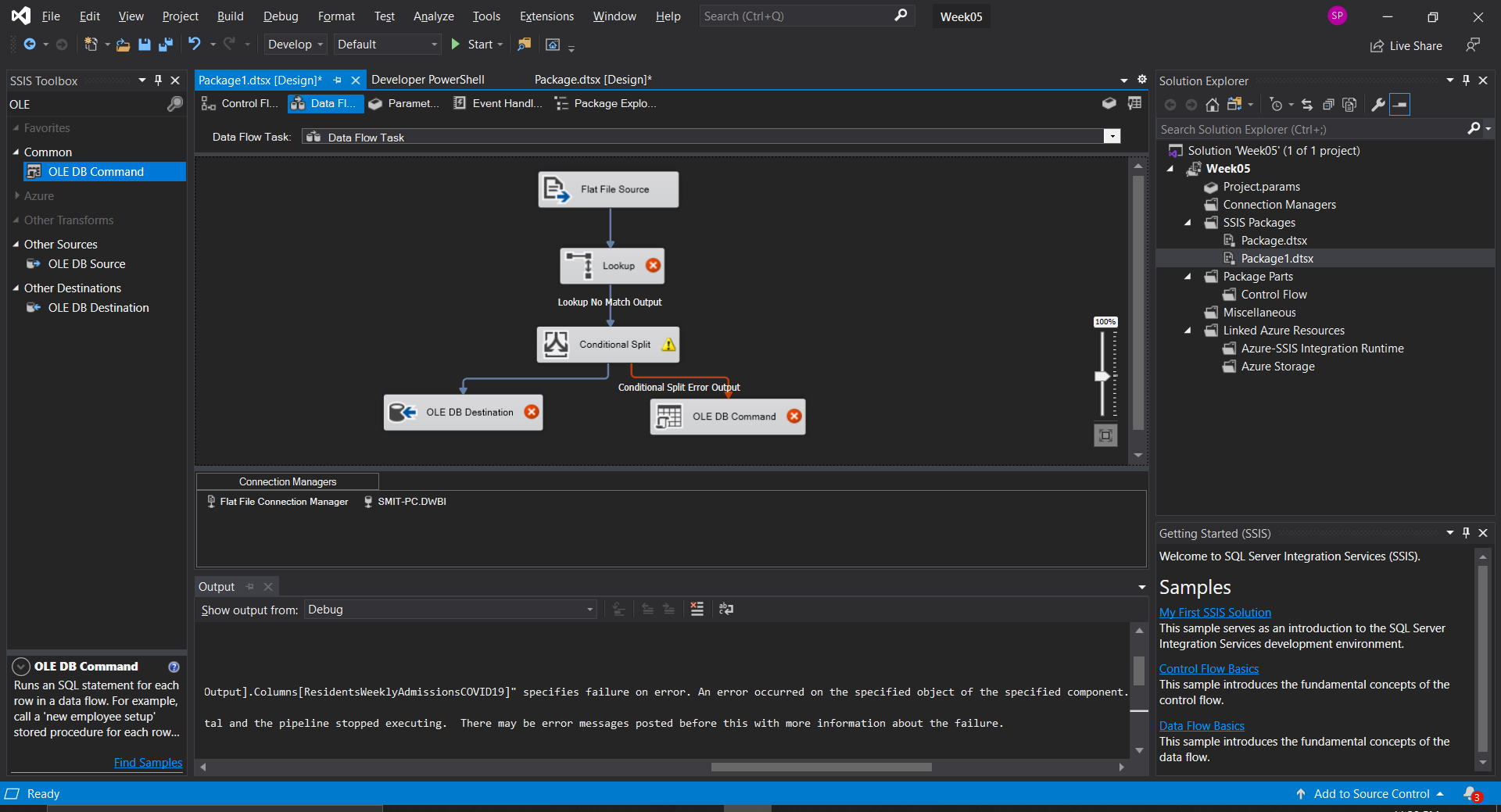
If we get the file again next month, we can load the data by using Incremental load process. In the Incremental Load is process of comparing the Target data table against the Source Data based on a Id Column or Date Stamp or Time Stamp and updating or loading the data to Target table as per below conditions:

* If there are any New records in Source data then we have to insert those records in Target data table.

For example, on a Daily/Weekly basis we have to insert the Region wise sales data.

* If there are any updated values in Source data then we have to update those records Target data table.

1. Modify the SSIS package to handle the next month of data



1. Create a new file of test data show how you created the input file

There are multiple ways one can create a test data file. We can either write a SQL DDL to CREATE a file in SQL itself or we can also hard code values in excel. There are various algorithms that can generate random numbers and variable to create a sample copy of the data. We have ensure that the data type of the test data matches to our original data. We can also create a test data file by using a subset of our original data. We can than change some variables of that data so that we get a unique set of data and data type constrains are also satisfied. I have used the later approach.

