MINI PROJECT 2

ASSIGNMENT

SUBMISSION REPORT

DATA MINING & WAREHOUSING

9/29/2016

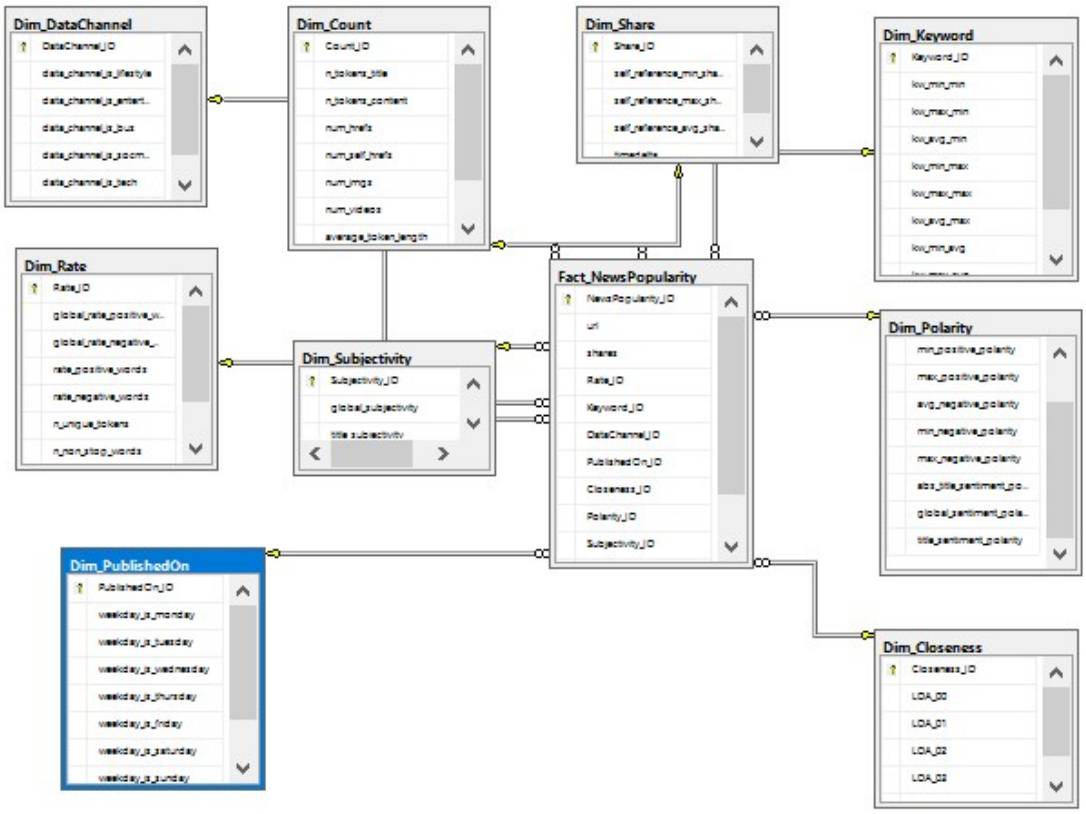
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
FACULTY OF COMPUTING

Gunathilaka D. D. T. M. (IT13011130) (WD-SE)
4th year 2nd semester 2016

The Following report describes the implementation procedure of data warehouse for "Online News Popularity" dataset. The data set summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. (www.mashable.com). Data set contains information relevant to each department of the company. Therefore, a data warehouse has been implemented to integrate all those information of articles for analysis and reporting purposes.

TABLE STRUCTURE

In order to create data warehouse **Star schema** design methodology has been used.



Facts table

The fact table contains measures that we take to evaluate the performance. In this case study the predicted number of shares measured given for various articles.

Dimension tables

The dimension tables contain different perspectives that we can take measurements from respective facts table. Dimensions are taken considering articles ratios, keywords, data channel, dates, closeness, polarity, counts and shares.

The Microsoft SQL Management Studio and MSSQL server 2012 are used to create the fact and dimensions tables.

In order to keep consistency and uniquely identify dimension table record Surrogate keys are declared for dimension tables.

```
Create Database OnlineNewsPopularity
Go
Use OnlineNewsPopularity
Go
```

Dimension tables

```
Drop Table Dim_Rate
Go
Create Table Dim_Rate
(
Rate_ID bigint identity(1,1) not null primary key nonclustered,
global_rate_positive_words real,
global_rate_negative_words real,
rate_positive_words real,
rate_negative_words real,
   n_unique_tokens real,
   n_non_stop_words real,
   n_non_stop_unique_tokens real
)
Go
Select * From Dim_Rate
```

```
Drop Table Dim_Keyword
Go
Create Table Dim_Keyword
(
Keyword_ID bigint identity(1,1) not null primary key nonclustered,
kw_min_min bigint,
kw_max_min bigint,
kw_avg_min real,
kw_min_max bigint,
kw_max_max bigint,
kw_avg_max real,
kw_min_avg real,
kw_max_avg real,
kw_avg_avg real
)
Go
Select * From Dim_Keyword

Drop Table Dim_DataChannel
Go
Create Table Dim_DataChannel
(
DataChannel_ID bigint identity(1,1) not null primary key nonclustered,
data_channel_is_lifestyle bigint,
data_channel_is_entertainment bigint,
data_channel_is_bus bigint,
data_channel_is_socmed bigint,
data_channel_is_tech bigint,
data_channel_is_world bigint
)
Go
Select * From Dim_DataChannel
```

```
Drop Table Dim_PublishedOn
Go
Create Table Dim_PublishedOn
(
PublishedOn_ID bigint identity(1,1) not null primary key nonclustered,
weekday_is_monday  bigint,
weekday_is_tuesday bigint,
weekday_is_wednesday bigint,
weekday_is_thursday bigint,
weekday_is_friday bigint,
weekday_is_saturday bigint,
weekday_is_sunday bigint,
is_weekend bigint
)
Go
Select * From Dim_PublishedOn

Drop Table Dim_Closeness
Go
Create Table Dim_Closeness
(
Closeness_ID bigint identity(1,1) not null primary key nonclustered,
LDA_00 real,
LDA_01 real,
LDA_02 real,
LDA_03 real,
LDA_04 real
)
Go
Select * From Dim_Closeness
```

```
Drop Table Dim_Polarity
Go
Create Table Dim_Polarity
(
Polarity_ID bigint identity(1,1) not null primary key nonclustered,
avg_positive_polarity real,
min_positive_polarity real,
max_positive_polarity real,
avg_negative_polarity real,
min_negative_polarity real,
max_negative_polarity real,
abs_title_sentiment_polarity real,
global_sentiment_polarity real,
title_sentiment_polarity real
)
Go
Select * From Dim_Polarity

Drop Table Dim_Subjectivity
Go
Create Table Dim_Subjectivity
(
Subjectivity_ID bigint identity(1,1) not null primary key nonclustered,
global_subjectivity real,
title_subjectivity real,
abs_title_subjectivity real
)
Go
Select * From Dim_Subjectivity
```

```
Drop Table Dim_Count
Go
Create Table Dim_Count
(
Count_ID bigint identity(1,1) not null primary key nonclustered,
n_tokens_title bigint,
n_tokens_content bigint,
num_hrefs bigint,
num_self_hrefs bigint,
num_imgs bigint,
num_videos bigint,
average_token_length real,
num_keywords bigint
)
Go
Select * From Dim_Count

Drop Table Dim_Share
Go
Create Table Dim_Share
(
Share_ID bigint identity(1,1) not null primary key nonclustered,
self_reference_min_shares bigint,
self_reference_max_shares bigint,
self_reference_avg_sharess real,
timedelta bigint,
)
Go
Select * From Dim_Share
```

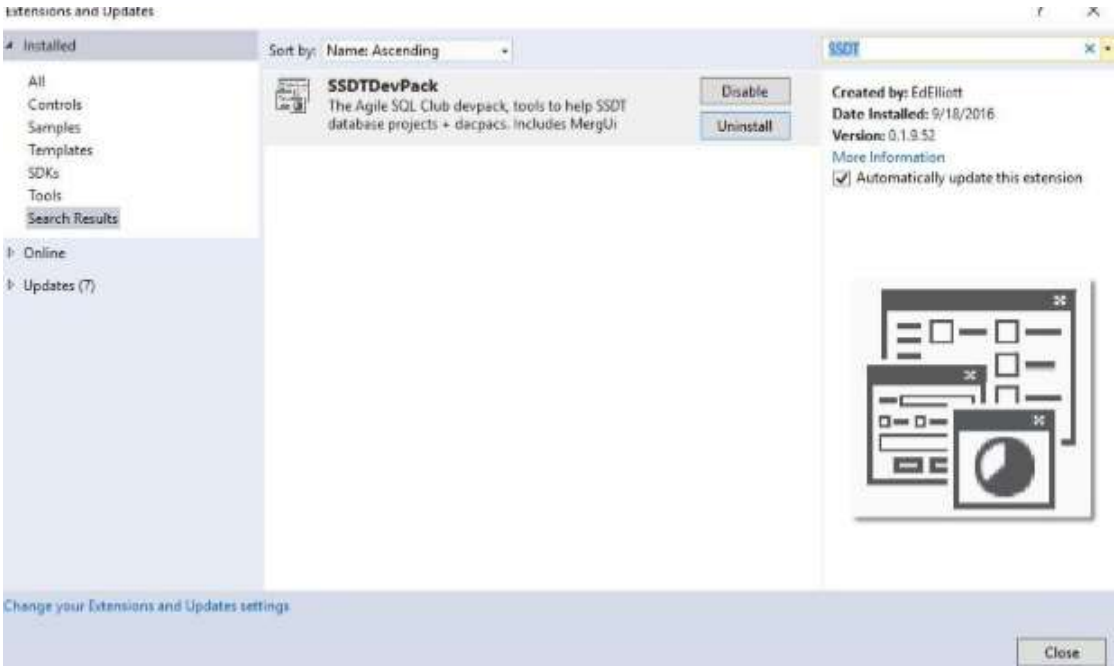
Facts table

```
Drop Table Fact_NewsPopularity
Go
Create Table Fact_NewsPopularity
(
url nvarchar(4000),
shares bigint,
Rate_ID bigint not null references Dim_Rate(Rate_ID),
Keyword_ID bigint not null references Dim_Keyword(Keyword_ID),
DataChannel_ID bigint not null references Dim_DataChannel(DataChannel_ID),
PublishedOn_ID bigint not null references Dim_PublishedOn(PublishedOn_ID),
Closeness_ID bigint not null references Dim_Closeness(Closeness_ID),
Polarity_ID bigint not null references Dim_Polarity(Polarity_ID),
Subjectivity_ID bigint not null references Dim_Subjectivity(Subjectivity_ID),
Count_ID bigint not null references Dim_Count(Count_ID),
Share_ID bigint not null references Dim_Share(Share_ID),
constraint NewsPopularity_pk primary key nonclustered
	(
		Rate_ID , Keyword_ID , DataChannel_ID , PublishedOn_ID ,
		Closeness_ID , Polarity_ID , Subjectivity_ID , Count_ID , Share_ID
	)
)
Go
Select * From Fact_NewsPopularity
```

TEMPLATE CONFIGURATION

The Visual Studio 2015 IDE is used to implement data integration and analysis tasks. However unlike Visual Studio 2013, 2012 or older versions Business Intelligence template comes as an extension to Visual Studio 2015. Therefore, we have to manually install it as an extension.

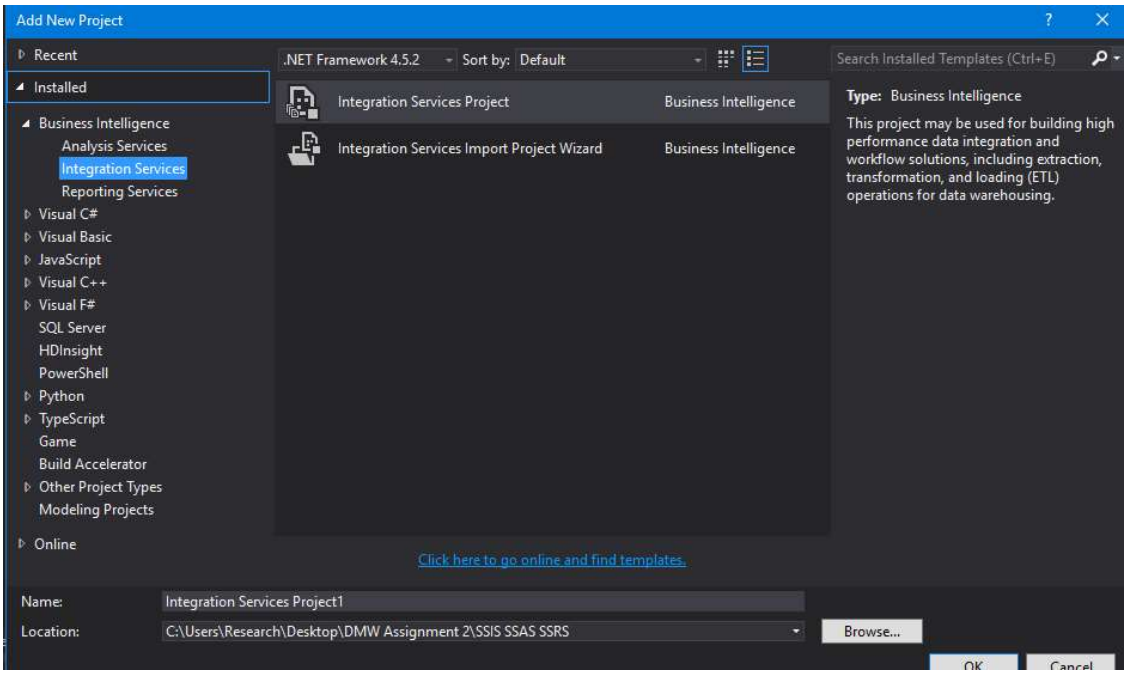
In Visual Studio 2015 IDE go to Tools → Extensions and Updates and in online tab (left) type "SSDT "(i.e. SSDT stands for SQL Server Data Tools) and click install "SSDTDevPack".



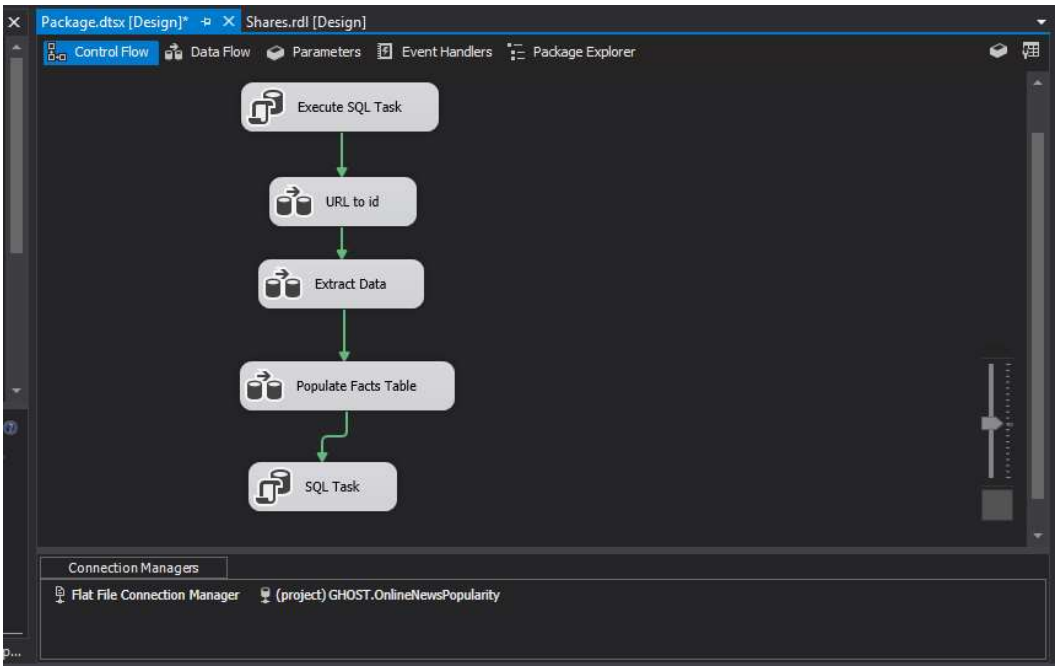
Once installation is completed restart IDE. Now we can create Business Intelligence projects in Visual Studio 2015. This contains sub project templates namely; "SQL Server Integration Service"," SQL Server Analysis Service" and "SQL Server reporting service".

SQL SERVER INTEGRATION SERVICES (SSIS)

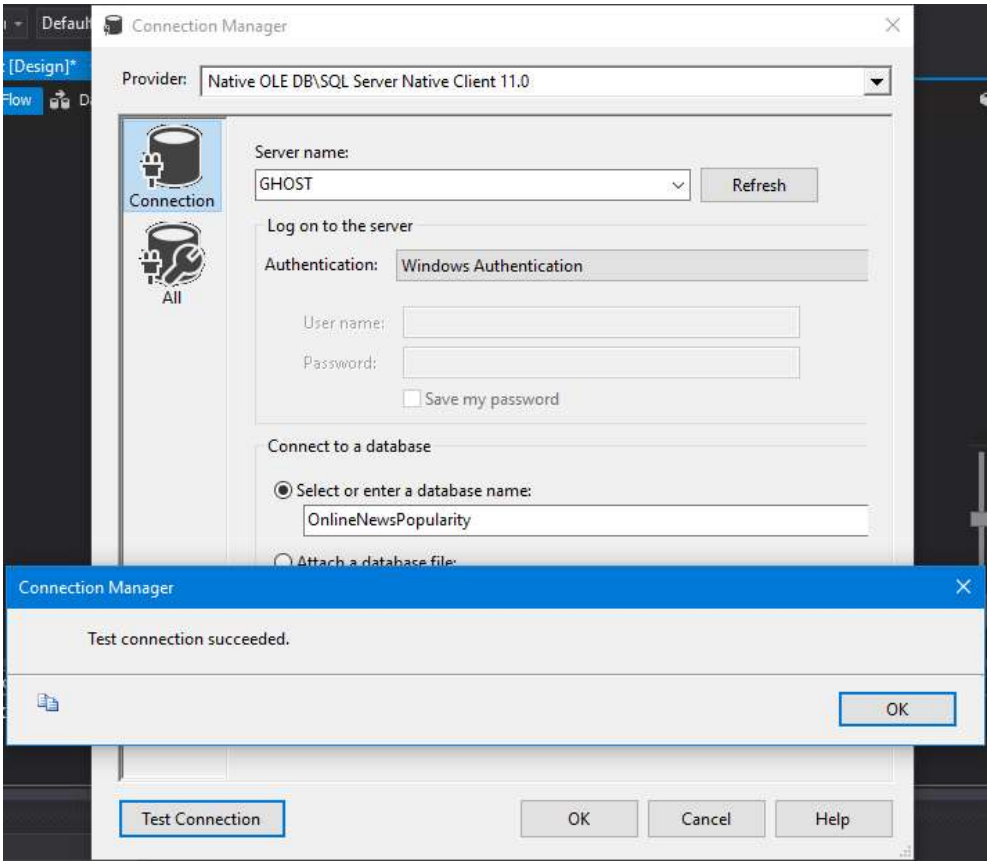
The integration services help to integrate the facts and dimension tables which origin from different sources in the hospital system. Hence firstly it has to implement an Integration service project in Visual Studio. (File> New Project >Installed > Business Intelligence>Integration Services Project)



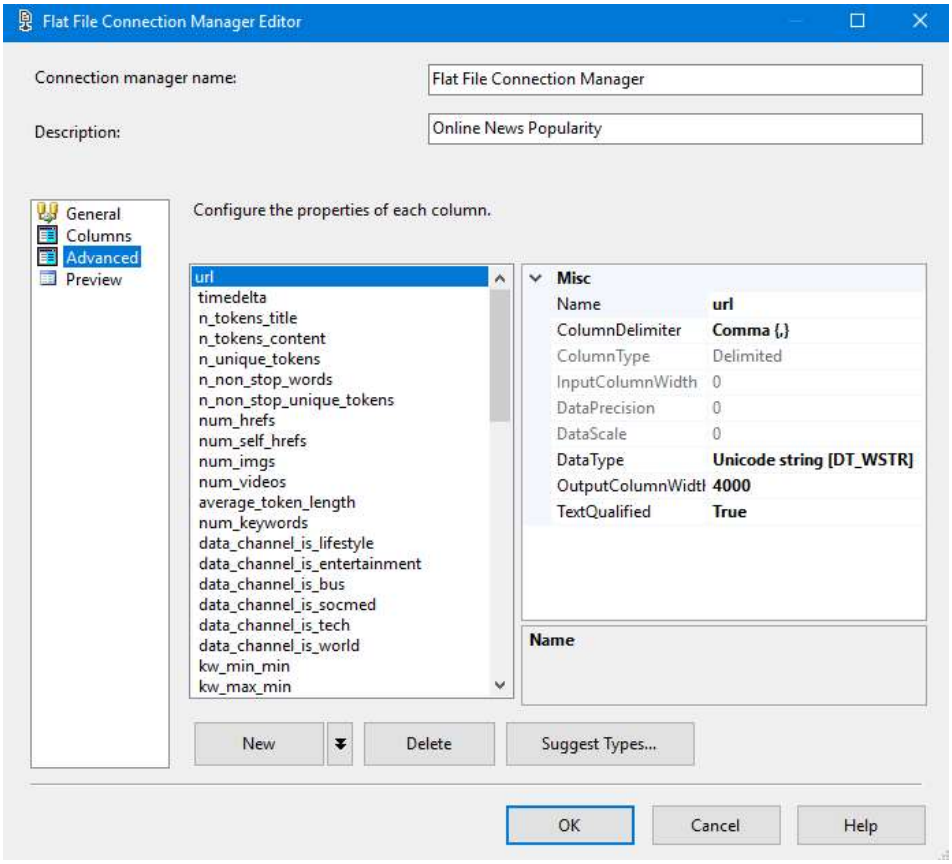
Once the project has been created add Data Flow task into Control Flow view.



OLEDB Database Connection with the MS SQL Server



Flat File Connection with Data Source



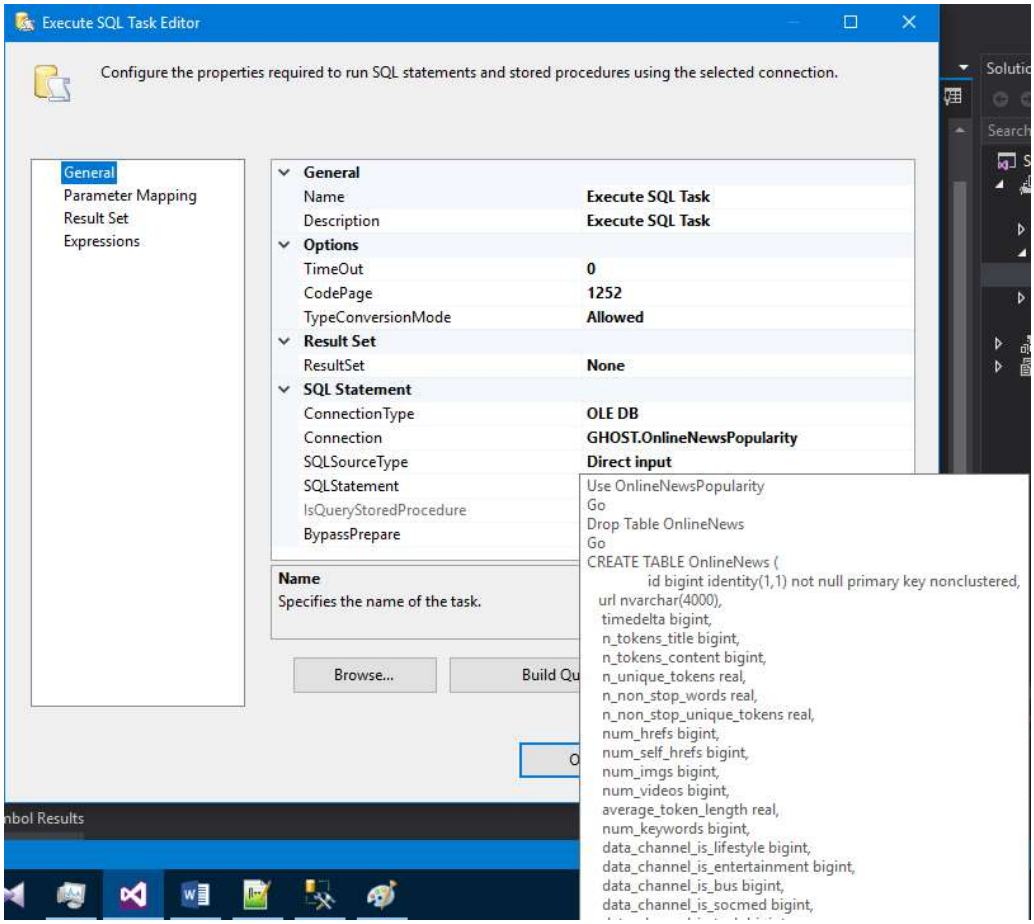
Column(Type)	Data Type
Url	Unicode String
kw_min_min	four-byte signed integer
Int columns	four-byte unsigned integer
Other	float

Implement integration Service (SSIS)

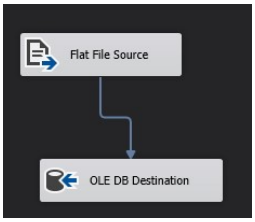
There are No missing of null data in the data set. Dataset was identified by the url column. Therefore, identity columns implemented.

Data inserted into a temporary transaction table and identity column generated for unique urls.

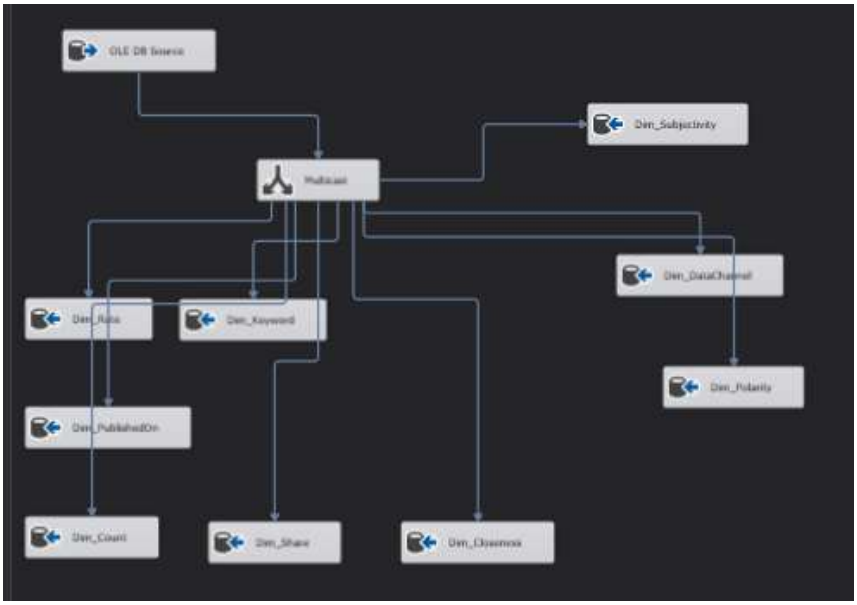
Create Temporary Table



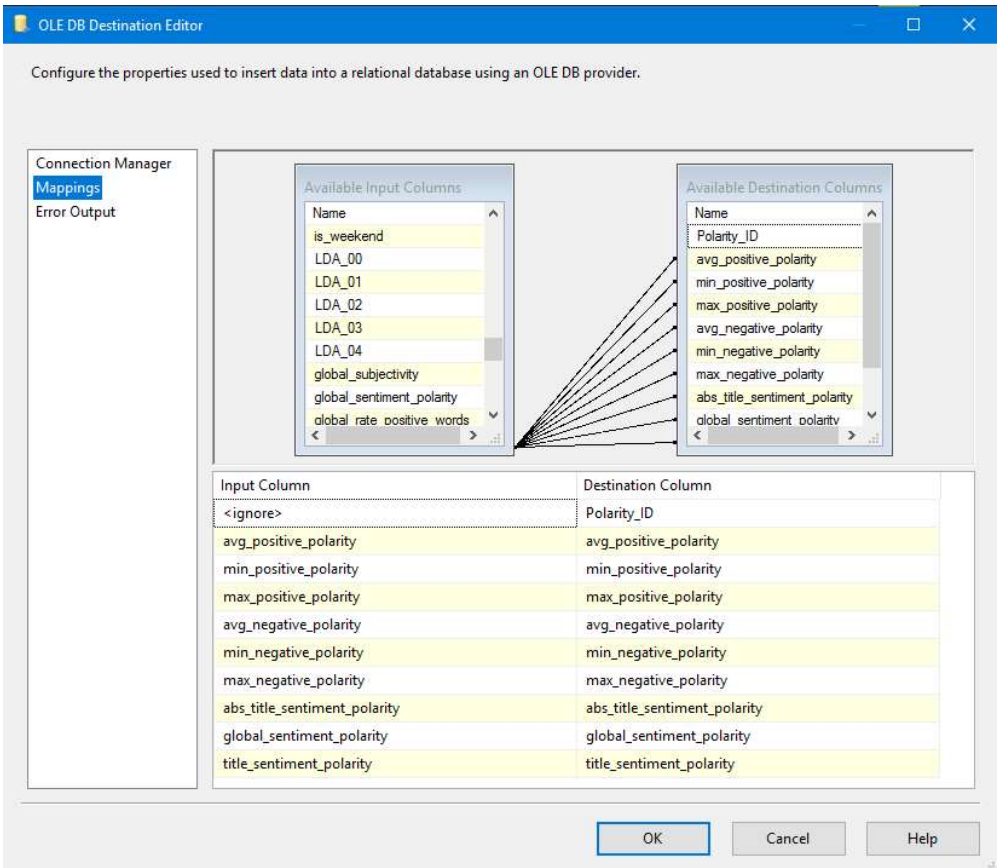
Url to ID



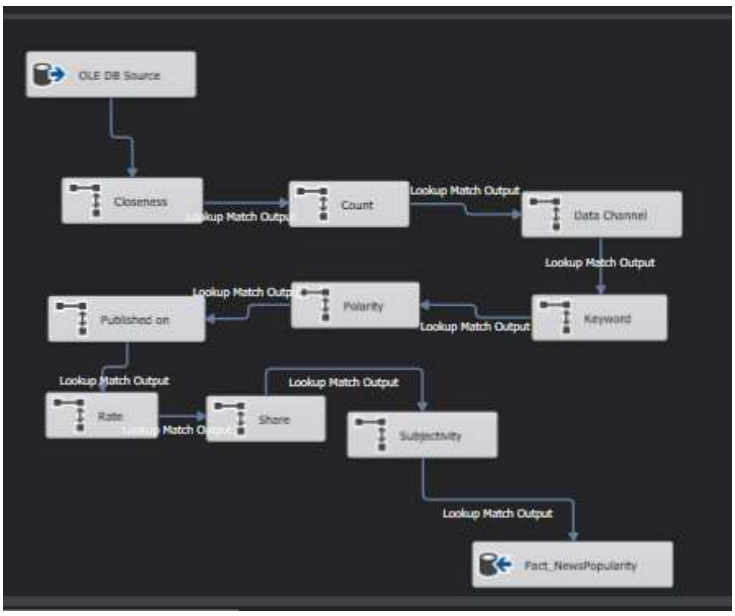
Extract Data



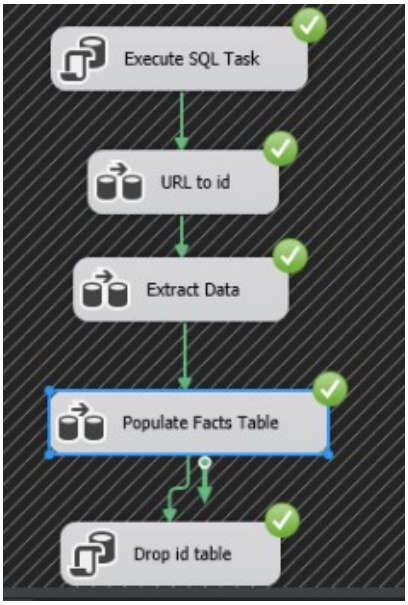
Each mapping is handled similar as below



Populate Facts Table



Lastly, the temporary table was dropped



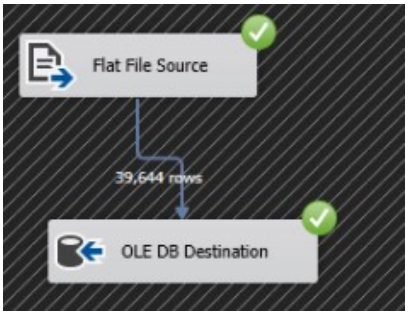


Figure 1: Url to ID

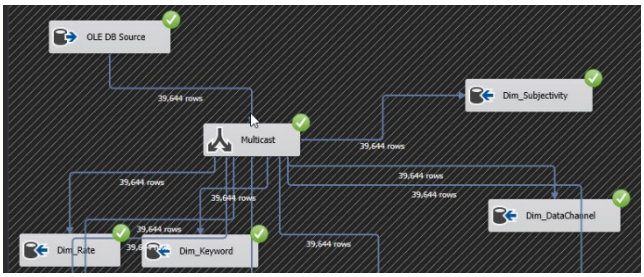


Figure 2 : Extract Data I

Figure 3 : Extract Data II

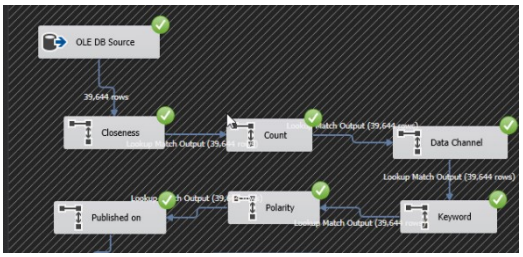


Figure 4 : Populate Fact Table I

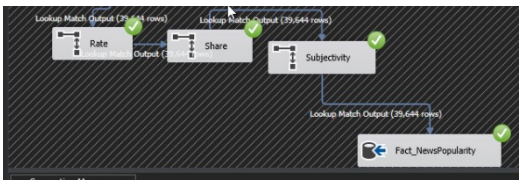
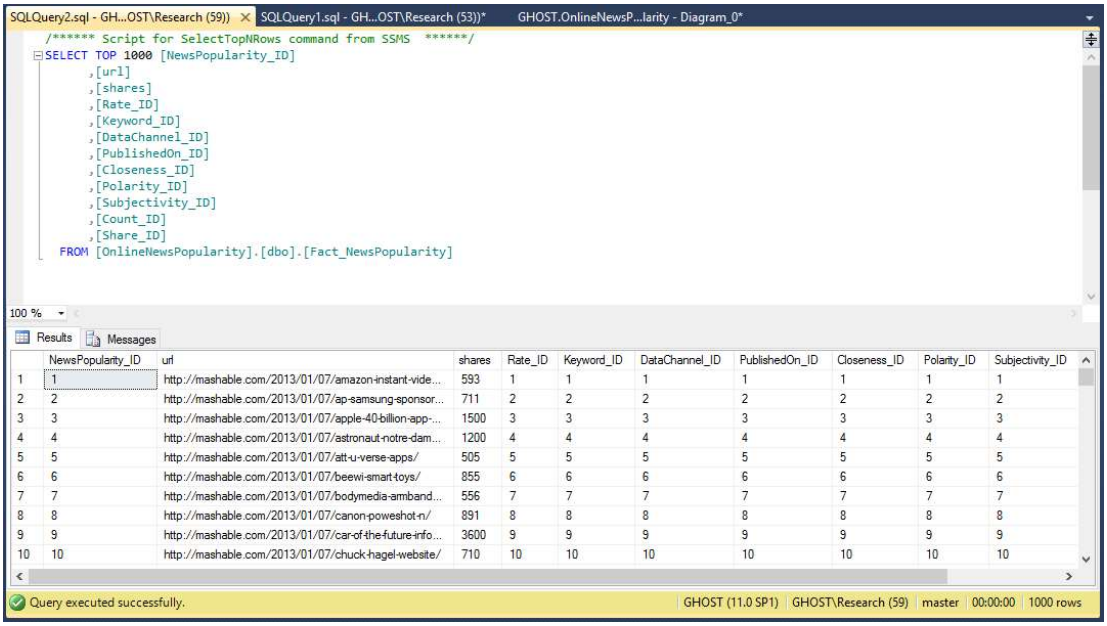


Figure 5 : Populate Fact Table II

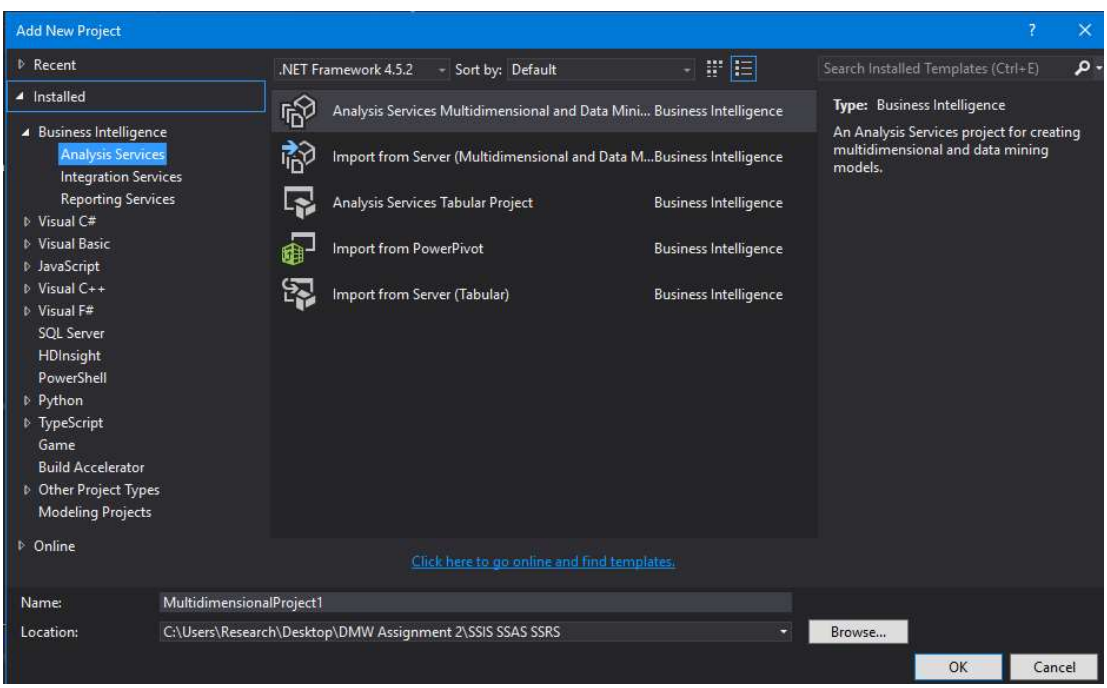
Inserted Values to Fact Table



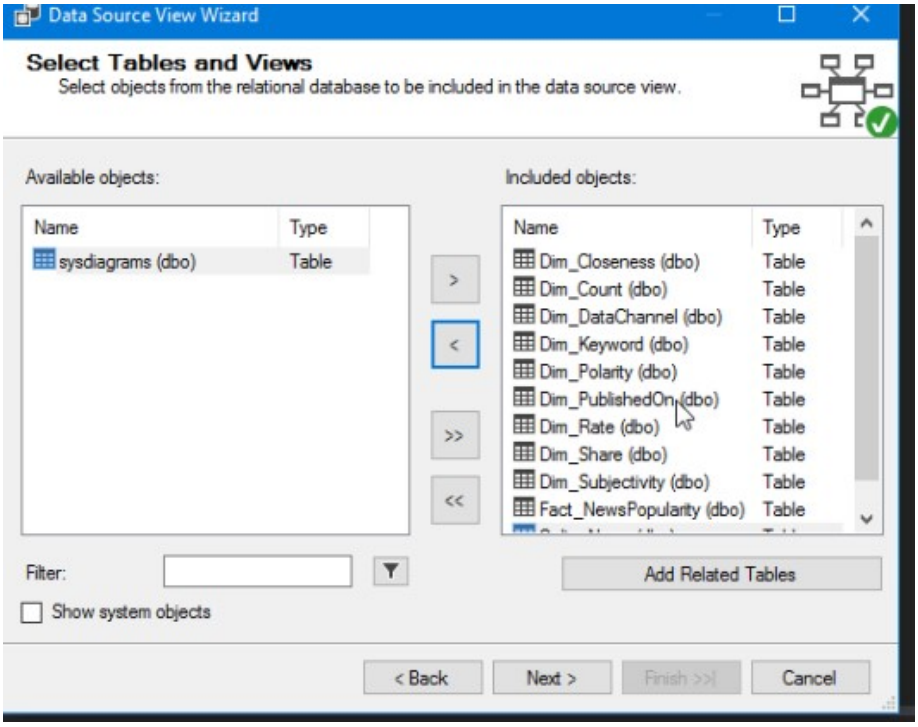
	NewsPopularity_ID	url	shares	Rate_ID	Keyword_ID	DataChannel_ID	PublishedOn_ID	Closeness_ID	Polarity_ID	Subjectivity_ID
1	1	http://mashable.com/2013/01/07/amazon-instant-vide...	593	1	1	1	1	1	1	1
2	2	http://mashable.com/2013/01/07/ap-samsung-sponsor...	711	2	2	2	2	2	2	2
3	3	http://mashable.com/2013/01/07/apple-40-billion-app-...	1500	3	3	3	3	3	3	3
4	4	http://mashable.com/2013/01/07/astronaut-notre-dam...	1200	4	4	4	4	4	4	4
5	5	http://mashable.com/2013/01/07/att-u-verse-apps/	505	5	5	5	5	5	5	5
6	6	http://mashable.com/2013/01/07/beewi-smart-toys/	855	6	6	6	6	6	6	6
7	7	http://mashable.com/2013/01/07/bodymedia-armband...	556	7	7	7	7	7	7	7
8	8	http://mashable.com/2013/01/07/canon-poweshot-n/	891	8	8	8	8	8	8	8
9	9	http://mashable.com/2013/01/07/car-of-the-future-info...	3600	9	9	9	9	9	9	9
10	10	http://mashable.com/2013/01/07/chuck-hagel-website/	710	10	10	10	10	10	10	10

Implement Analysis Services (SSAS)

Create new analysis project



Create New Data Source with Dimensions and Fact Table



Created Schema

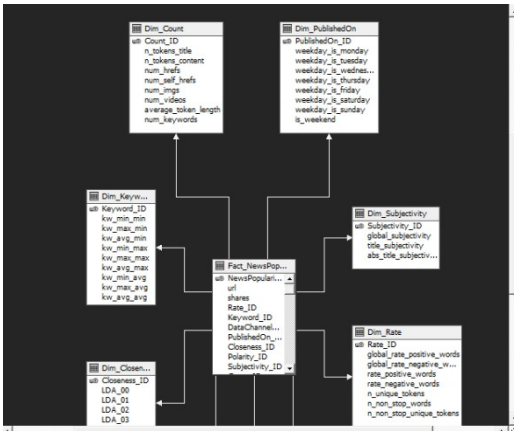


Figure 6 : Creates Schema I

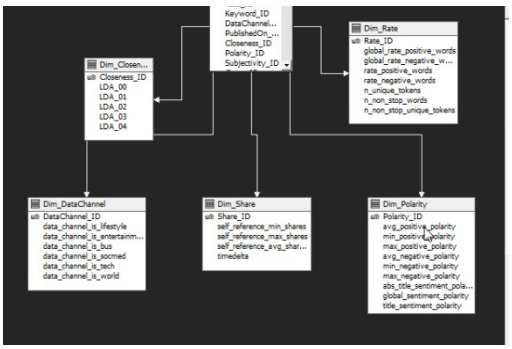
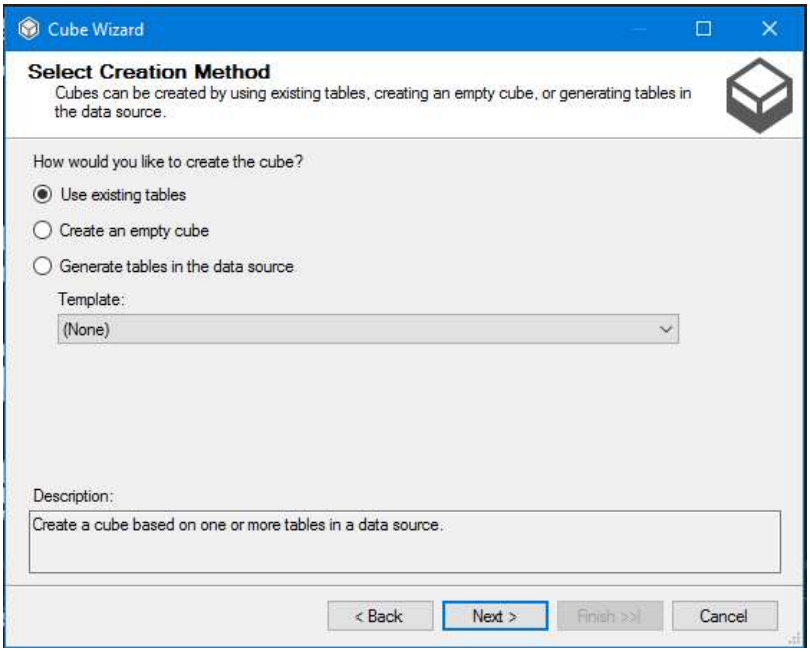
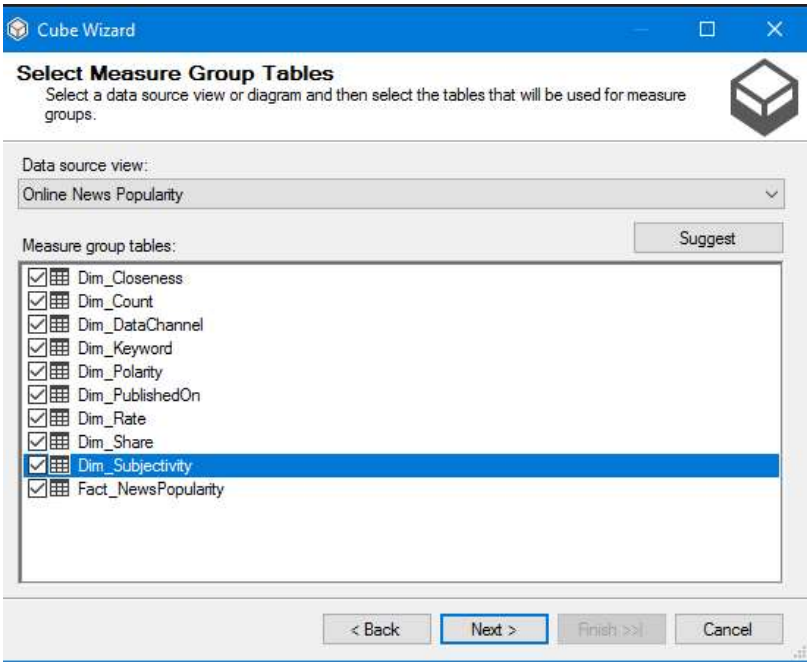
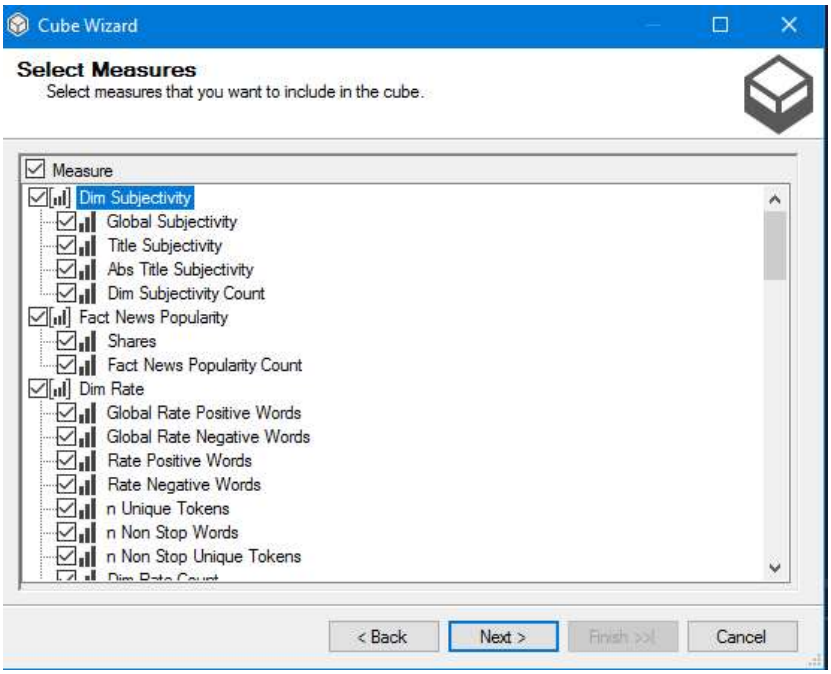


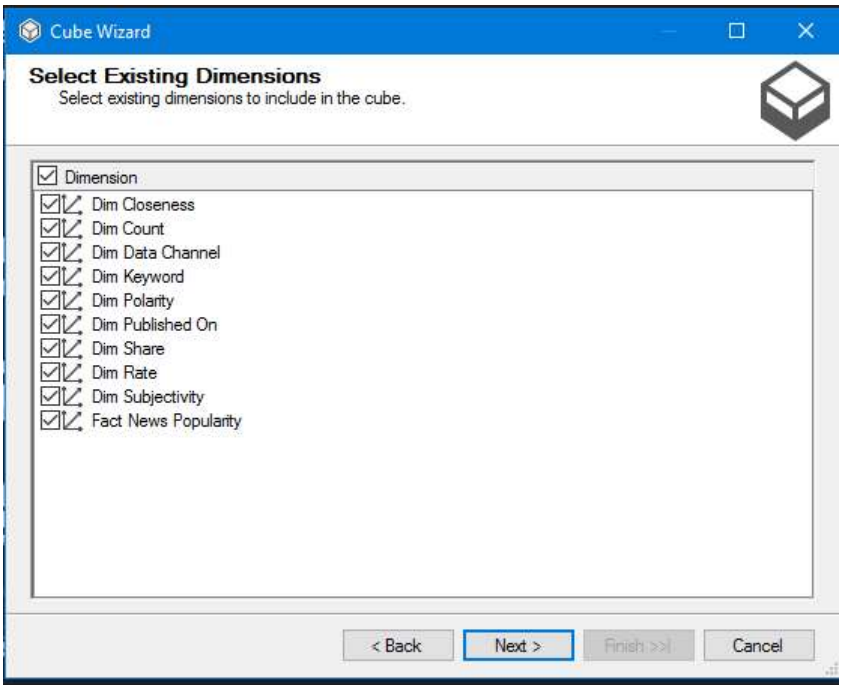
Figure 7 : Created Schema II

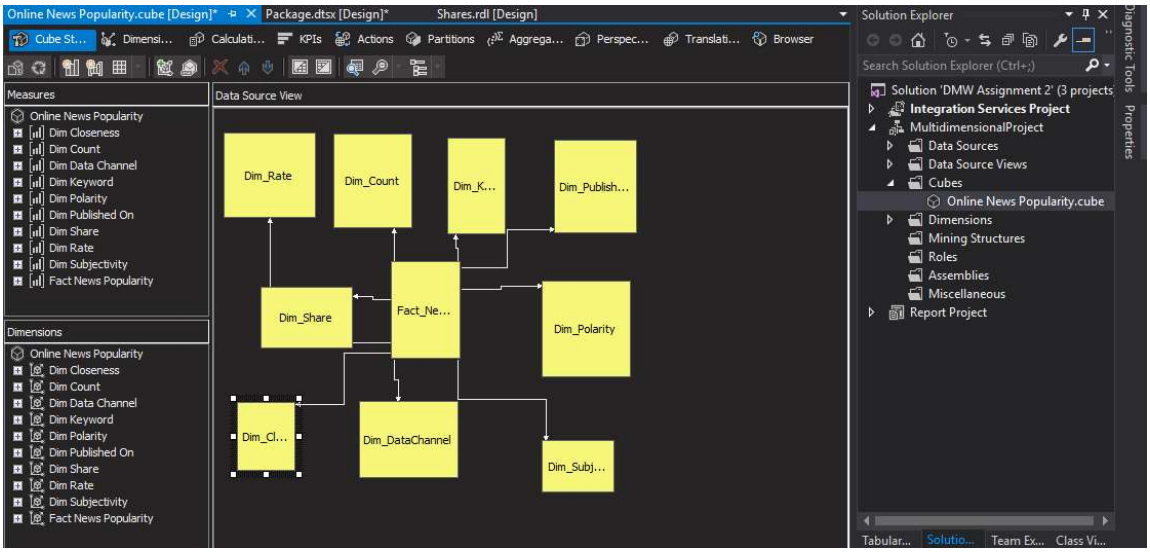
Create New Cube



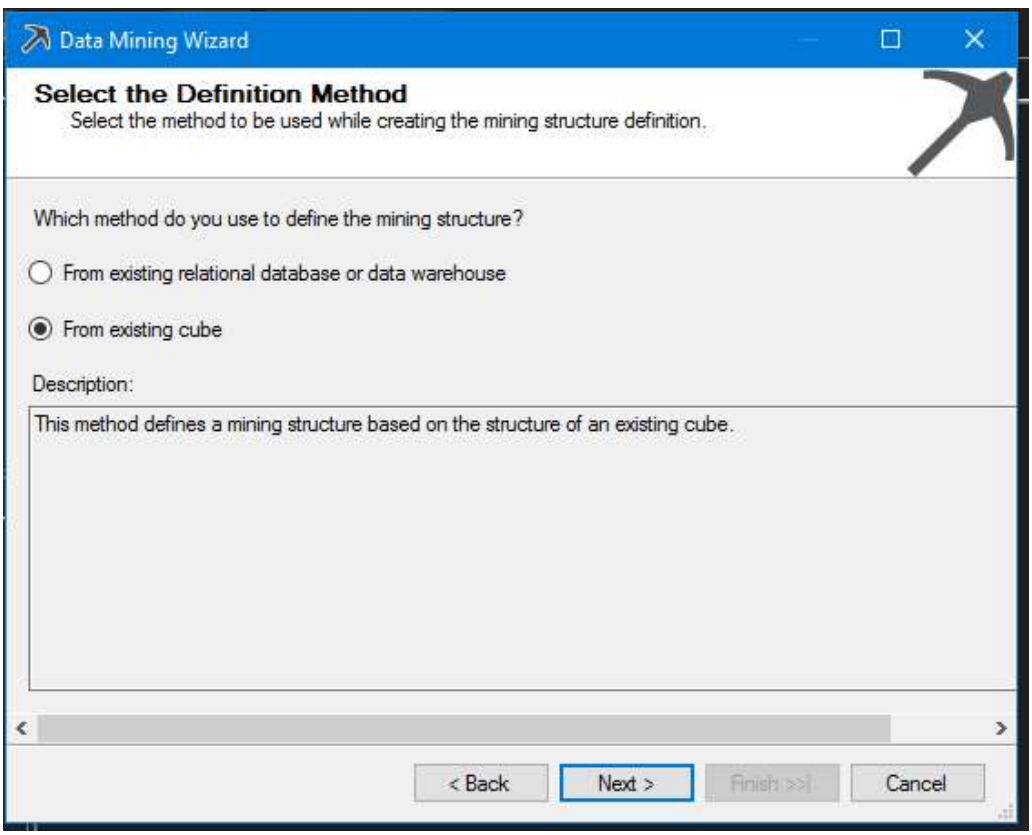


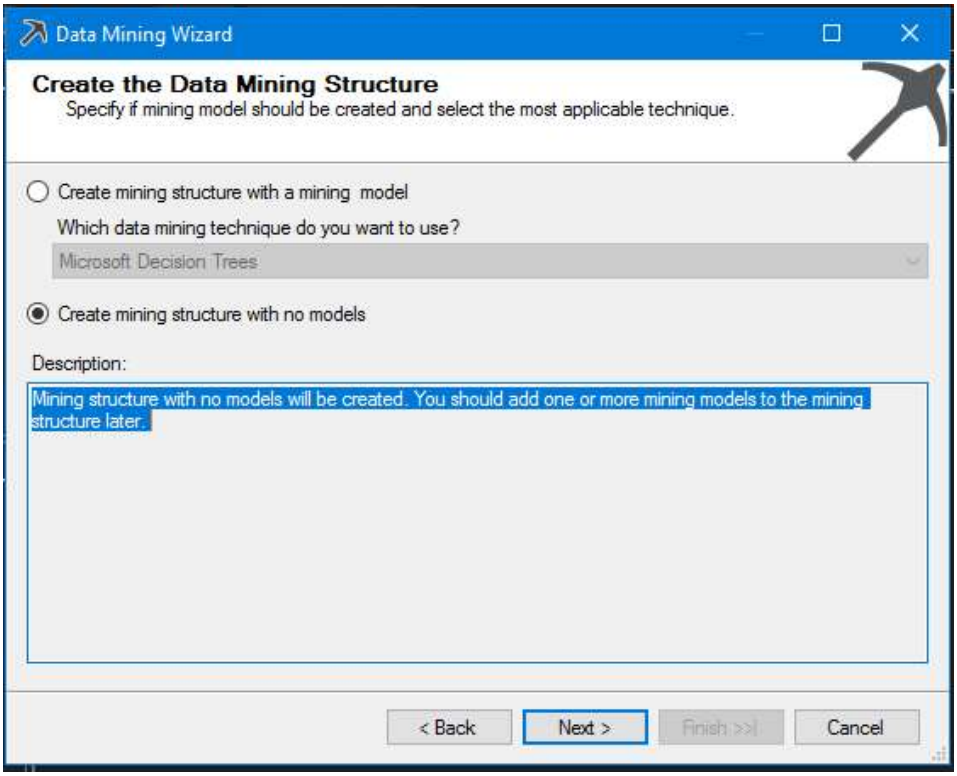

Cube Wizard
Select Measures
Select measures that you want to include in the cube.
Measure
Dim Subjectivity
Global Subjectivity
Title Subjectivity
Abs Title Subjectivity
Dim Subjectivity Count
Fact News Popularity
Shares
Fact News Popularity Count
Dim Rate
Global Rate Positive Words
Global Rate Negative Words
Rate Positive Words
Rate Negative Words
n Unique Tokens
n Non Stop Words
n Non Stop Unique Tokens
< Back
Next >
Finish >>|
Cancel

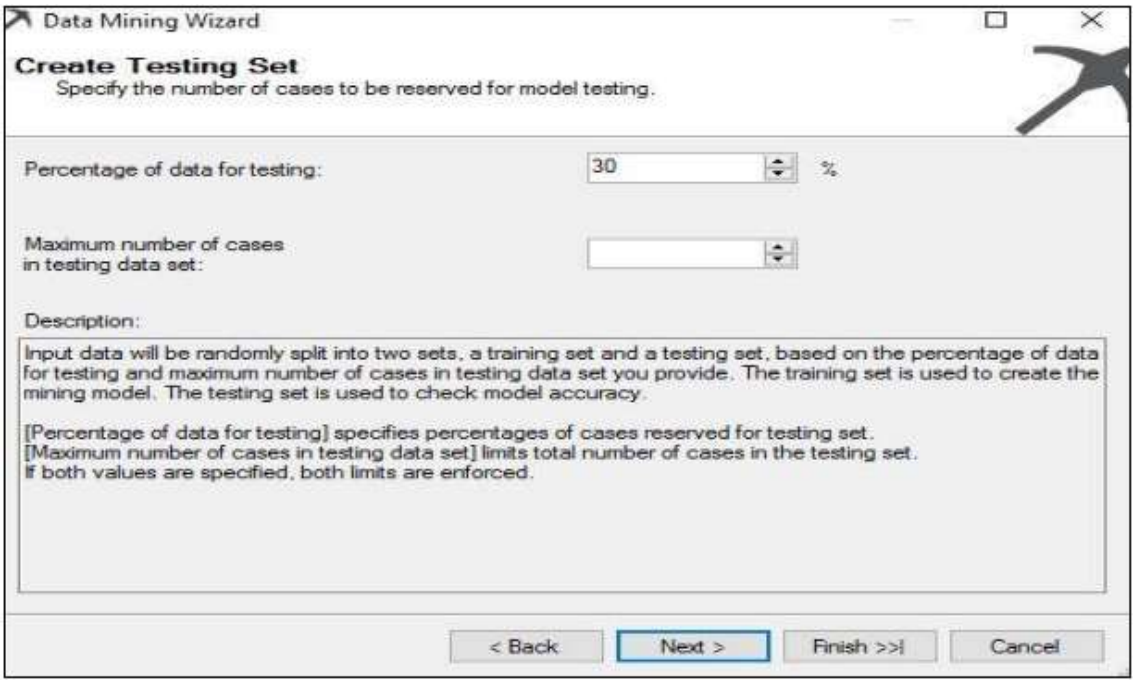

Cube Wizard
Select Existing Dimensions
Select existing dimensions to include in the cube.
Dimension
Dim Closeness
Dim Count
Dim Data Channel
Dim Keyword
Dim Polarity
Dim Published On
Dim Share
Dim Rate
Dim Subjectivity
Fact News Popularity
< Back
Next >
Finish >>|
Cancel



Create Mining Model

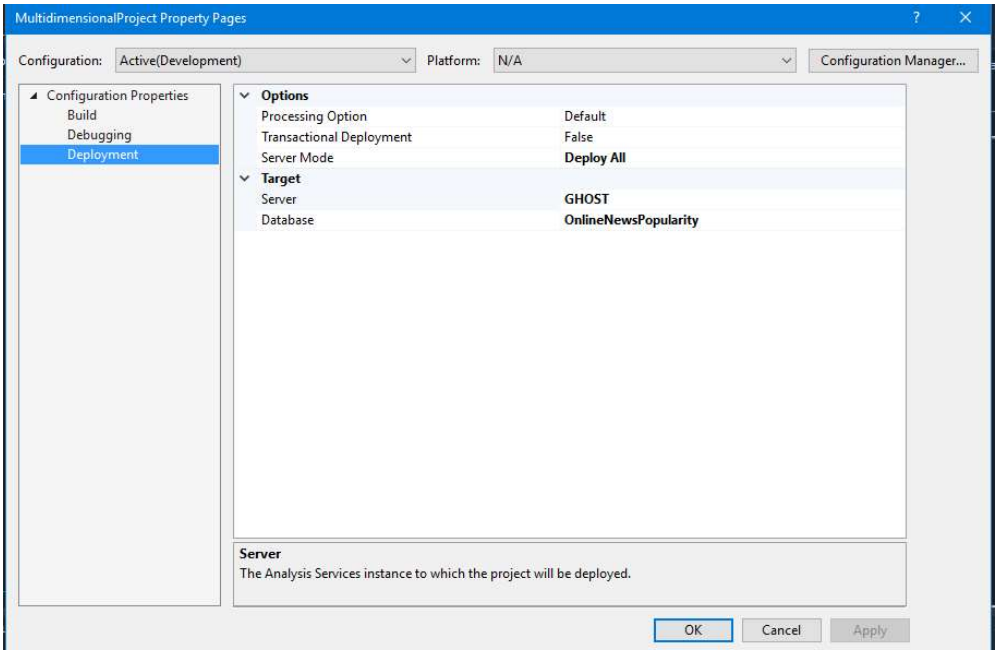


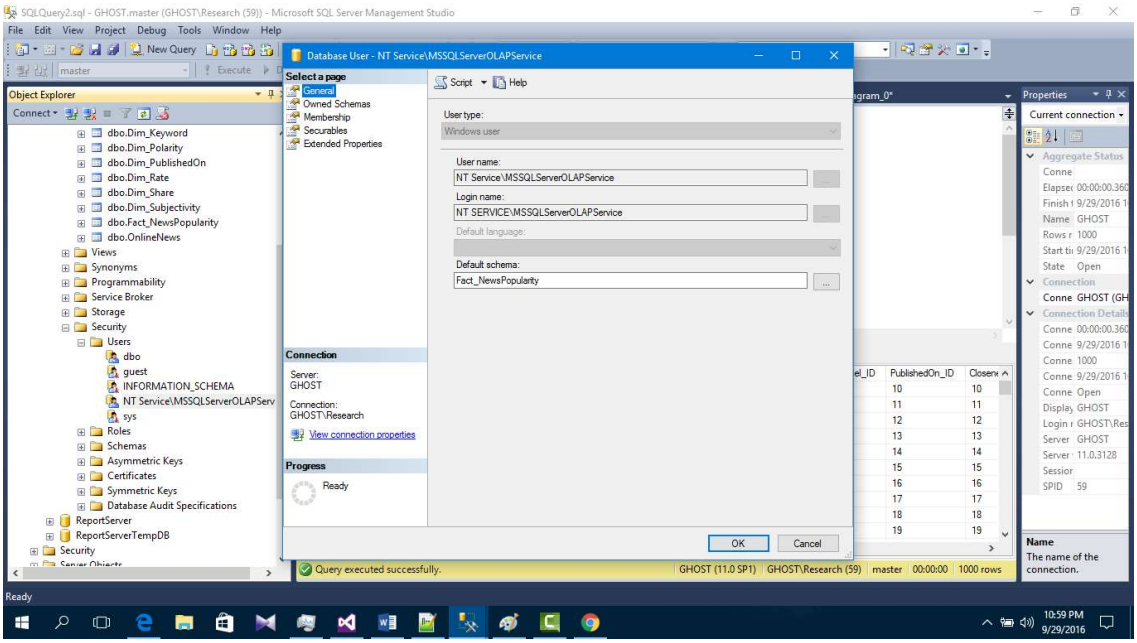

Data Mining Wizard
Create the Data Mining Structure
Specify if mining model should be created and select the most applicable technique.
Create mining structure with a mining model
Which data mining technique do you want to use?
Microsoft Decision Trees
Create mining structure with no models
Description:
Mining structure with no models will be created. You should add one or more mining models to the mining structure later.
< Back
Next >
Finish >>|
Cancel


Data Mining Wizard
Create Testing Set
Specify the number of cases to be reserved for model testing.
Percentage of data for testing:
30
%
Maximum number of cases in testing data set:
Description:
Input data will be randomly split into two sets, a training set and a testing set, based on the percentage of data for testing and maximum number of cases in testing data set you provide. The training set is used to create the mining model. The testing set is used to check model accuracy.
[Percentage of data for testing] specifies percentages of cases reserved for testing set.
[Maximum number of cases in testing data set] limits total number of cases in the testing set.
If both values are specified, both limits are enforced.
< Back
Next >
Finish >>|
Cancel

Click Finish

Deployment Settings

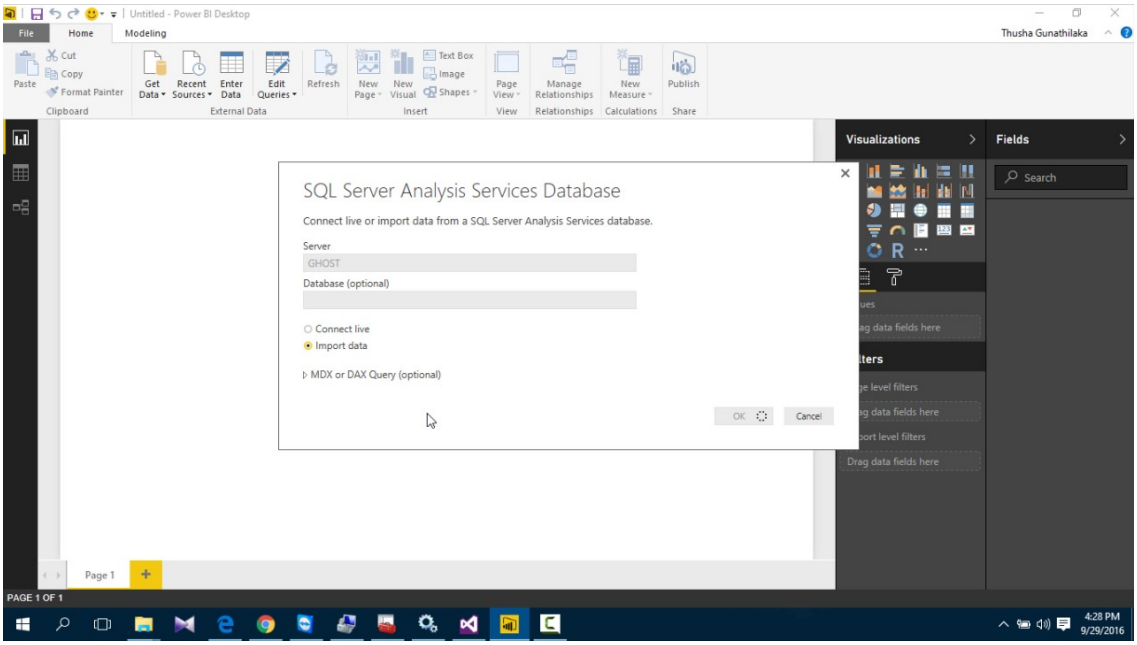


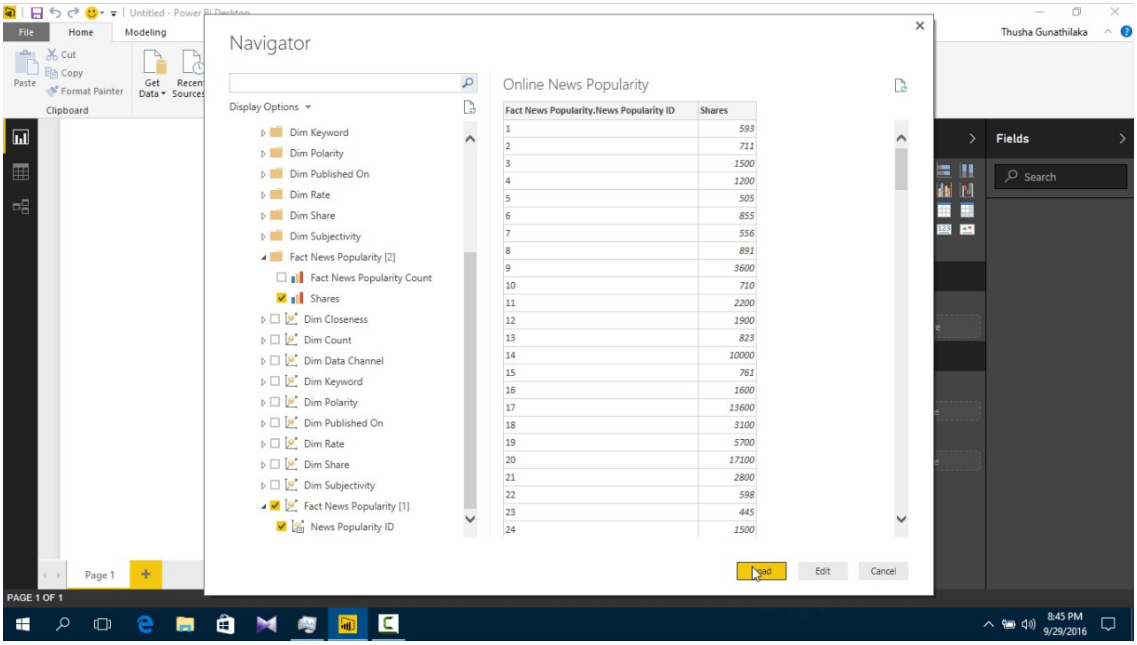


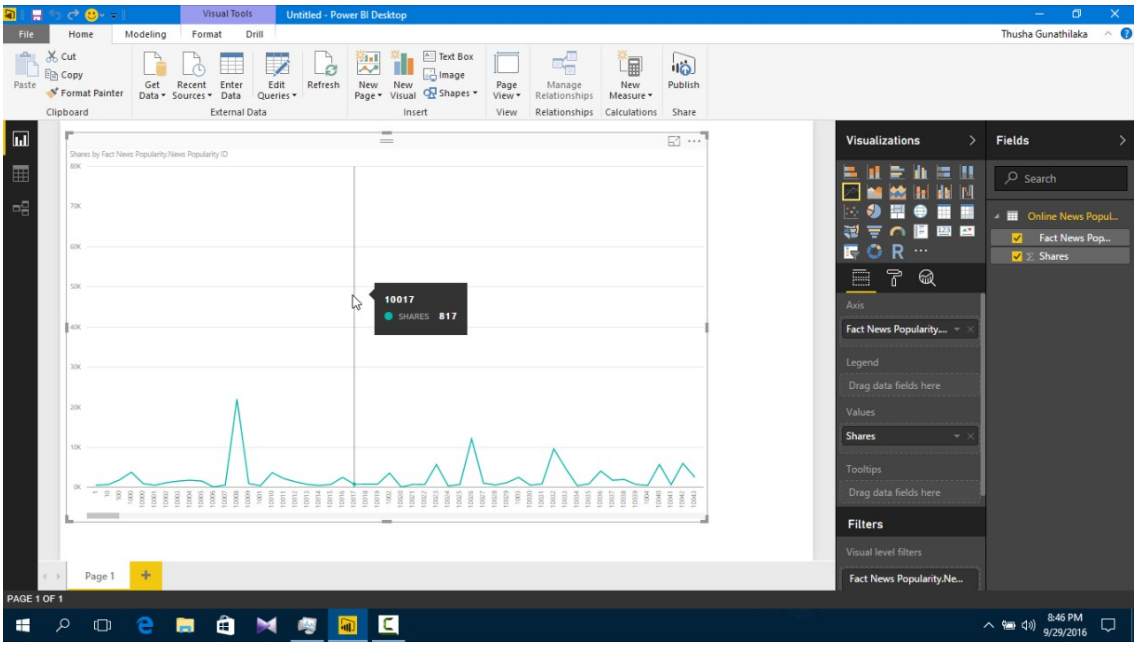
Set SSAS as Startup File

Run Solution

Visualizing Data with Power BI

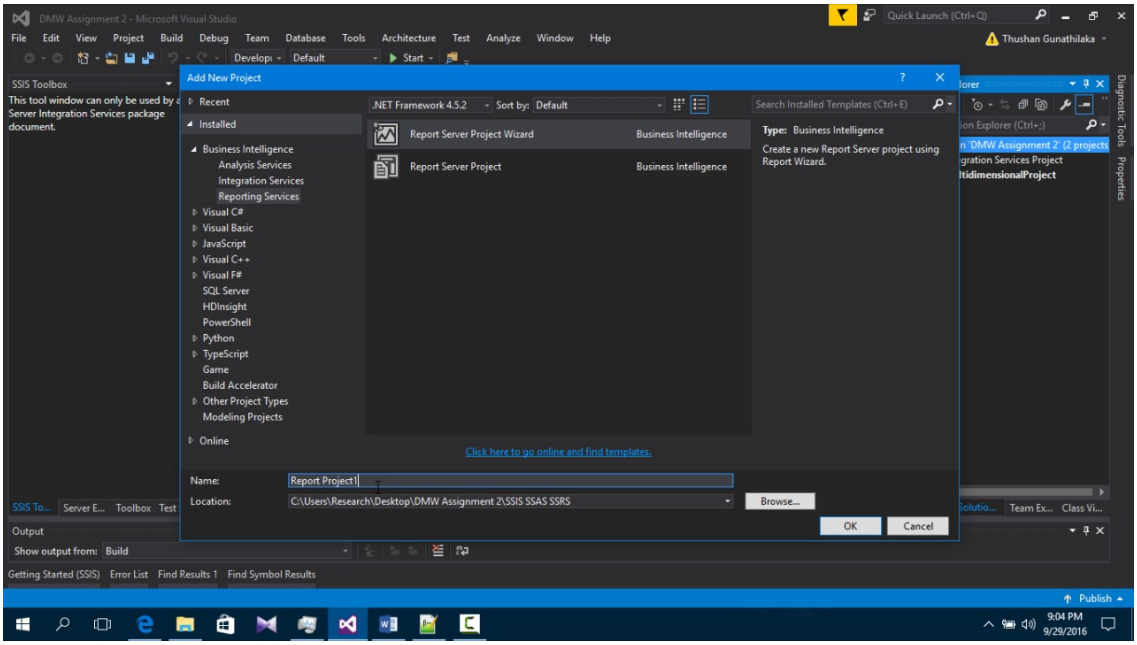




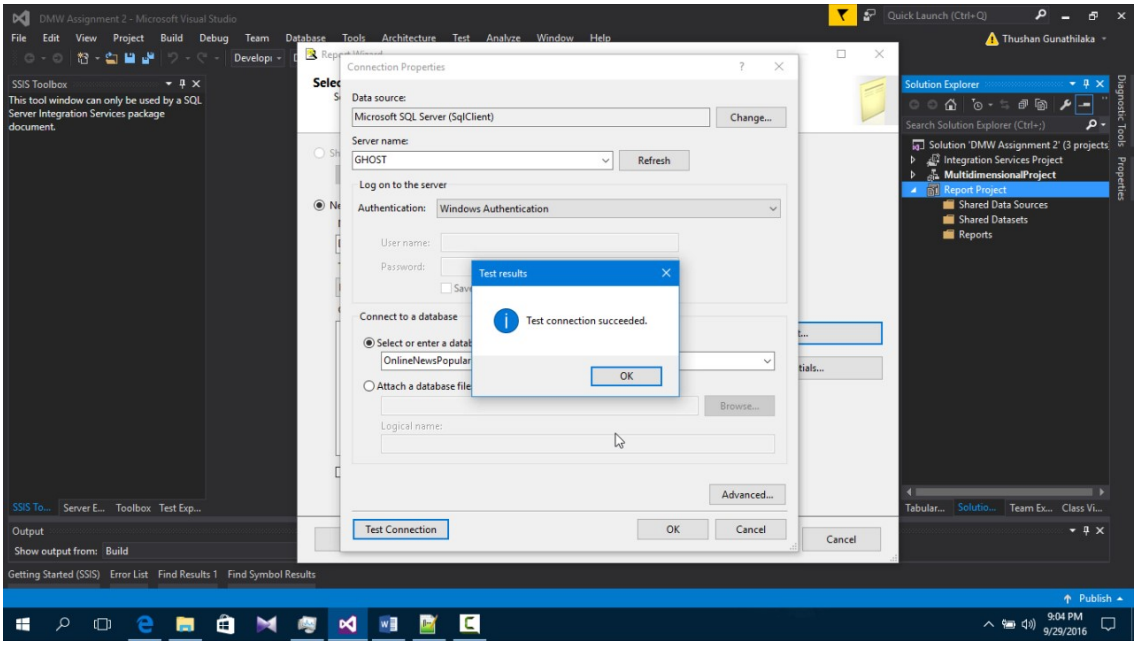


SQL SERVER REPORTING SERVICES (SSRS)

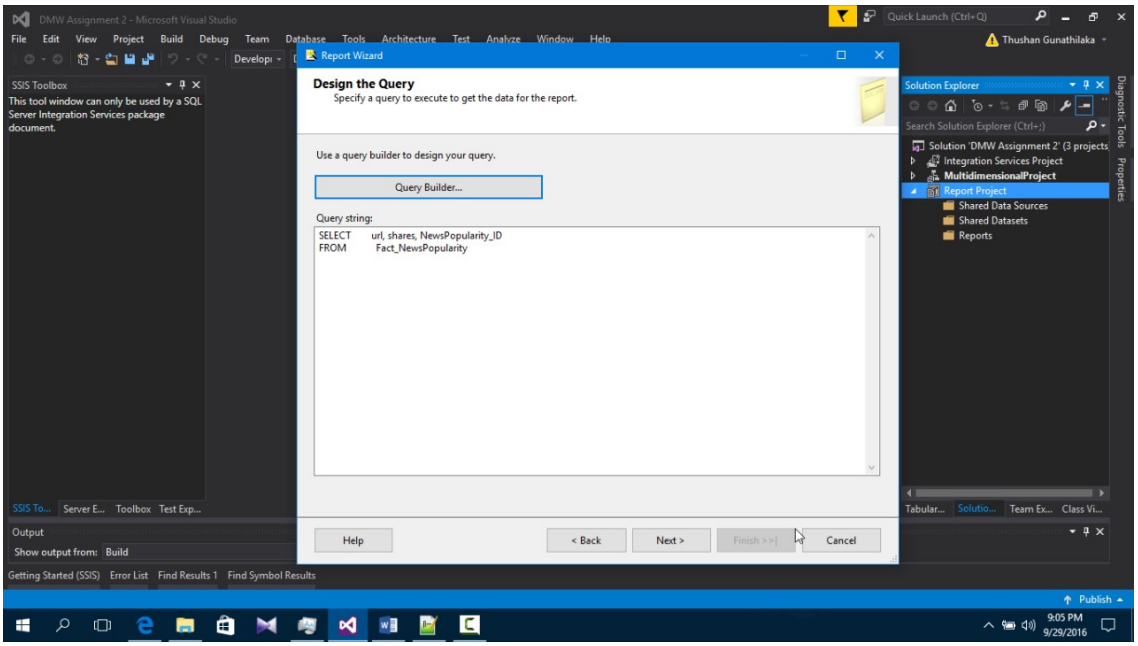
Create Reporting Services project in Business Intelligence template



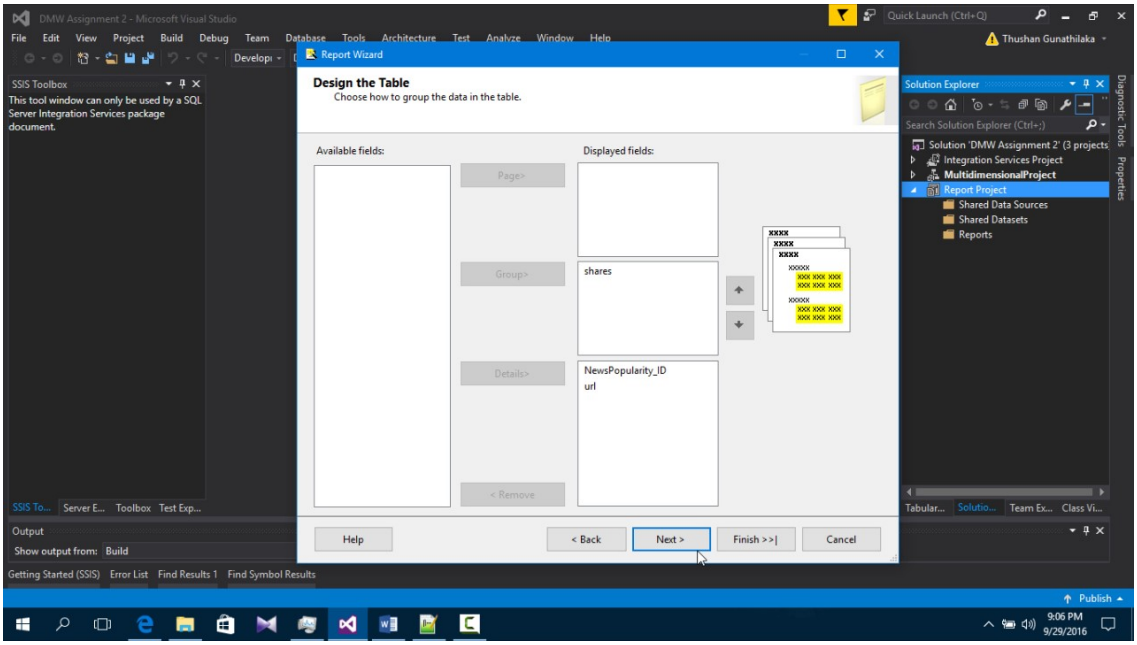
Set Data Source

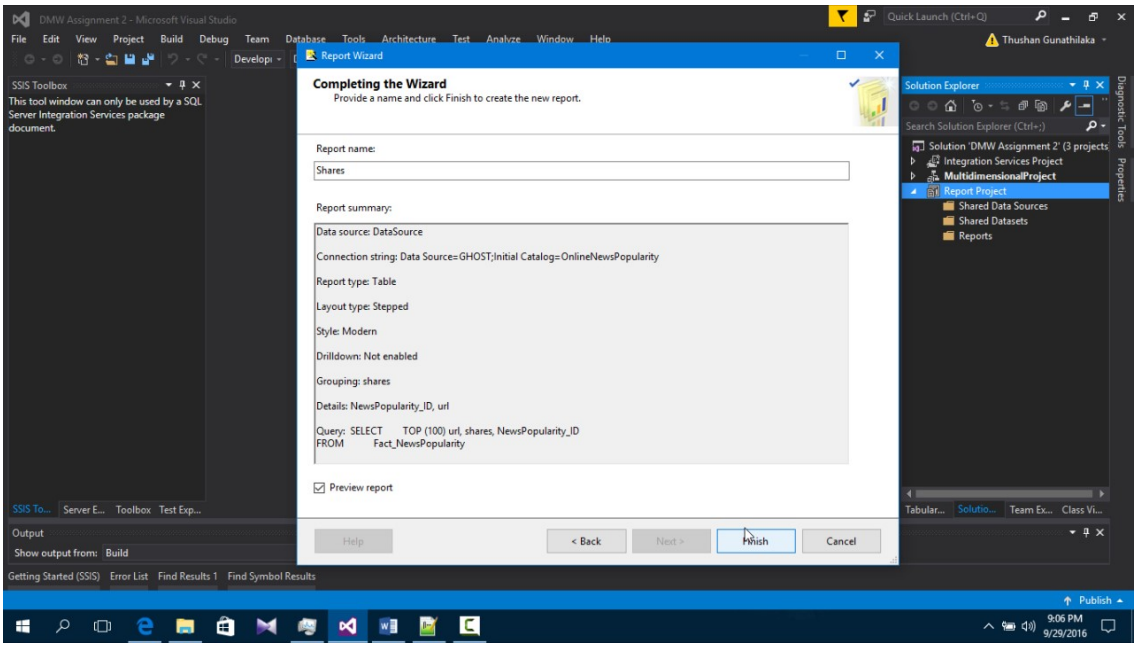


Extract Data



Tabular Report Sample





Sample Preview