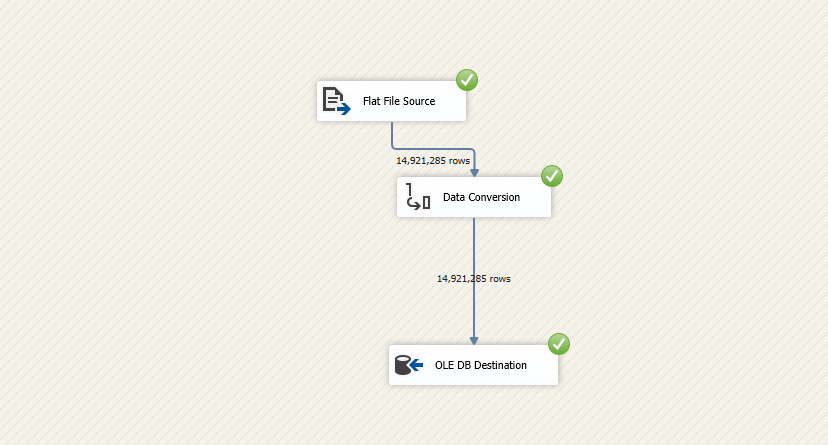
1. SSIS INJESTION OF THE FILE:
2. Approach to load the files: The files have been loaded using flat file source and data

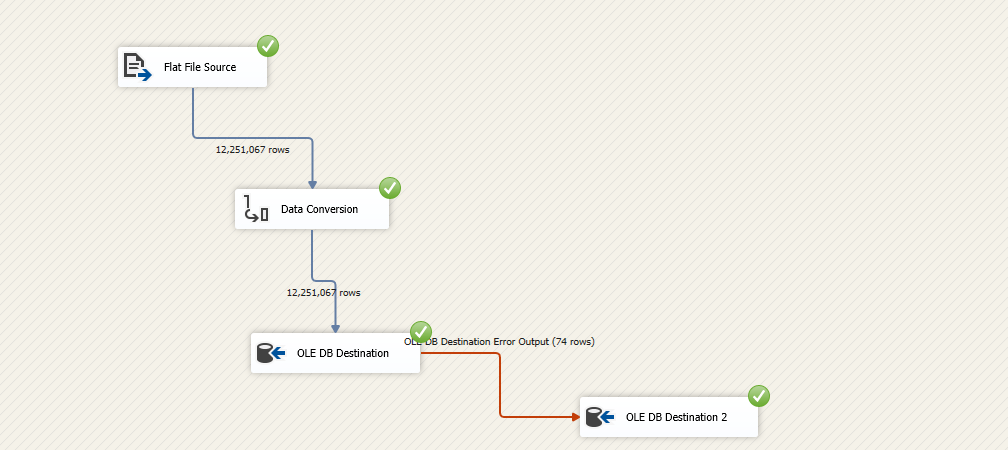
conversion to load the files into the SQL Server.

Issues Encountered while loading Queries.txt:



1. I encountered some issues with respect to memory which was resolved after running the visual studio using administrator account.
2. My initial testing was by taking a sample 100 records and loading them to the table which went fine as the QUERY column length was less. It was literally impossible to open the file as it was more than a GB so I was not aware of the length the column can go to.
3. Once I loaded the whole file I got error for truncation as the QUERY field was going more than what DT\_STR can hold.
4. So, I converted the field first to DT\_TEXT and then using data\_conversion transformation to DT\_STR which allowed me to load the data successfully.

Issues Encountered while loading Clicks.txt:



1. My initial testing was by taking a sample 100 records and loading them to the table which went fine as the QUERY column length was less. It was literally impossible to open the file as it was more than a GB so I was not aware of the length the column can go to.
2. Once I loaded the whole file I got error for truncation as the QUERY field was going more than what DT\_STR can hold.
3. So, I converted the field first to DT\_TEXT and then using data\_conversion transformation to DT\_STR which allowed me to load the data successfully.
4. However, there were 74 records that were giving error which were separately taken out into another table

2) Normalization Model: The model has been normalized to 3rd normal form.

Conceptual Model:

In the conceptual model, I have outlined the basic entities which I believe should be part of a portal system like Webjet. These entities have not been found in the clicks and query data set that I have downloaded.



Logical Model:



Cardinality and Optionality:

Cardinality:

1. As the logical diagram represent there are **1:M** relationship between User and User\_Transact table as when each user logs into the system he generates a user\_id and a query\_id. The User\_Transact table can have information about single user\_id’s having multiple query id’s hence 1:M relationship.
2. There is a **1:1** relationship between USER\_TRANSACT and QUERY table as of each query that is being searched a new Query\_id is generated which is always unique irrespective of the user. For the same user, also we will always have different Query\_id for each query. This query information is stored inside QUERY\_INFO table.
3. There is a **1:M** relationship between QUERY\_INFO and CLICKS\_INFO as for each query we get multiple links on the page. And for each click that is made on the link we get a click id. So, for multiple clicks we have same query\_id associated with the same.
4. There is a 1:M relationship between USER\_TRANSACT and CLICKS\_INFO as for each query\_id we have correspoinding multiple query\_id’s in the CLICK\_INFO table.

Optionality:

1. A user who logs in might just log in and come out without searching or putting any query. Hence there is a partial constraint between USER and USER\_TRANSACT.
2. A person can just query or search without logging into the system and hence there will be a query\_id generated without a user\_id and hence there is an optionality between users and query. A query may or may not have a user\_id associated with it.
3. Similarly, there will be an optional relation between QUERY\_INFO and CLICKS\_INFO as there will always be a click\_id associated with a query\_id but this is not true other way.

Assumptions:

1. There are certain structural assumptions that have been made where-in its assumed that a new entity USER will be present and other entities will fall in place with the ER -DIAGRAM. I have just considered the entities with respect to the data set and not the whole picture.
2. Certain Data assumptions have been made that query\_id will always be unique for each query that is fired by a user. The click\_id generated from each query is unique.

Data Governance:

1. Proper access to be given to users to log into the systems.
2. The storage must be maintained properly for such huge data coming in every day.
3. The data is central to the portal for its future analysis and must be protected using backups and proper privileges and roles in place.
4. Proper training to be given to new employees with respect to data architecture to ensure that data is appropriately taken care of and is very important for future application development and Misunderstood data or incomplete data requirements can affect the successful outcome of any Information Systems project, making the creation and maintenance of the organization’s Enterprise Data Model a truly beneficial exercise.
5. Dimensional Model:

The dimensional modelling that will be used here with the available information is the star schema based on the information present in the data sets.



There are 3 dimensions TIME, USER, QUERY and the fact table CLICKS which contains he number of clicks.

The thought process that went behind creating this dimensional model is to enhance faster retrieval for reporting purposes. The reports that might be useful for Business intelligence are the reports based on TIME, USER\_ID.

You can create Time based reports to generate number of clicks for month, quarter, year. You can create reports for User\_id to show which user has the most number of clicks.

Assumptions:

1. The model is created based on the available set of data in the two data sets. In a portal there might be more entities, which have not been considered here.
2. The model has been created with Business Intelligence perspective .

Data Governance:

1. Proper access to be given to users to log into the systems.
2. The storage must be maintained properly for such huge data coming in every day.
3. The data is central to the portal for its future analysis and must be protected using backups and proper privileges and roles in place.
4. Load detailed atomic data into dimension structures.
5. Ensure that every fact table has an associated dimension table.
6. Continuously balance requirements and realities to deliver a DW/BI solution that’s accepted by business users and that supports their decision-making.
7. All facts in a single fact table should be at the same grain or level of detail.
8. Dimensional models should be made around business processes.