

Data Warehouse Implementation for UK and IOWA accidents



Submitted to: Simon Caton

Student Name: Divyang Jain
Student ID: x16110323
Course: MSc in Data Analytics
National College of Ireland,
Dublin

CONTENTS

1. Introduction.....	3
2.Implementation and Architecture.....	6
3.Datawarehouse data model.....	7
4.ETL strategy.....	10
4.1 Extract	
4.2 Transform	
4.3 Load	
5.Applications of data warehouse(Reporting).....	15
6.References.....	20

1.Introduction

In this century of big data environment, data is growing day by day which is creating a lot of trouble for the devices to store the data. Data gets outdated(vanished) from the system and new data comes in but if a company wants to make a decision on the statistics of previous processes done which is vanished is a big challenge, so for that purpose we need a data warehouse to store the large amount of data. It stores the enormous data from many different sources created in past and the current time. Now mostly all types of business firms are heading towards the business intelligence and that can be done only by using data warehouse. Business intelligence helps the organization to analyze the operational data on daily bases which can be a good advantage for the company as it will improve the strategies where they lack behind.

Data in data warehouse is non-volatile and subject oriented which helps the organization in good decision making (Inmon, W.,2005). Sometimes centralized store data can create many difficulties as there will be a chain of data connected in which one fault will affect the other data which will create a chaos in the system. For this difficulty, new method is introduced in which we can create many distinct data marts for various areas which will eventually help in making good business intelligence reports by dividing it into various parts.

In this project, I am analyzing the data on accidents happened in UK and IOWA. Accidents are nowadays increasing day by day due to foggy weather, rainy weather or for many different reasons which involves many injuries and fatalities happened in different areas. I am analyzing the data to know on what **measures** government should take steps to reduce the increase of accidents by saving the lives of people. I have seen many accidents in my life and always wanted to minimize the effect of it by taking some measures, so I am building the data warehouse on accidents to give the right information which can be useful for government to take major steps to minimize it.

I have built the data warehouse to store the historical data of accidents and queries by integrating business intelligence which will convert data into information. It will tell the government like which area is most accident prone, in which weather, lightning conditions most injuries and fatalities happened and many queries which are covered later in this project which can be helpful for government to take measures on.

Collection of datasets:

For a reliable data warehouse, we need a good quality of data sources which includes proper format and without any wrong values.

In this project, I have used 3 data sets from three various sources two are structured, and one is unstructured.

Data Source 1 (Structured data): I have used official website of US state-IOWA which contains the legal information of everything happened in that state and recorded accordingly. I have taken the crash dataset from this site which is my first data set.US has maximum number of accidents in the world so there will

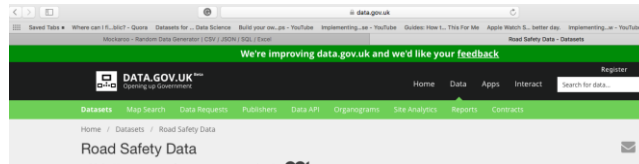
be a huge difference if I would have taken whole US data and UK data so I have taken IOWA that is one of the biggest state in US. It contains data like age,sex ,weather,light, and for much information go to this URL.



URL- <https://data.iowa.gov/Transportation-Utilities/Crash-Data/bew5-k5dr>

Data Source 2 (Structured): I have used another official website of UK which contains data of UK road accident specifically. It contains all the information of accidents happened in UK this motivated me to take dataset of road accidents from this site.

It is structured data which contains road accidents data like weather conditions, lightning conditions, injuries fatalities, number of vehicles and for more information go to this URL.



URL: <https://data.gov.uk/dataset/road-traffic-accidents>

Data Source 3 (Unstructured): I have taken the third data source which is a social networking website viz. Twitter because it's a famous social site from where we can get the information on public opinion. I have fetched tweets on different places (UK and IOWA) and made simple sentimental analysis which gives idea of which place has more number of positive and negative views (Matloff, N., 2011). I have done this on the basis of negative words and positive words dictionary. Code for fetching this information is below-

URL-(<https://gist.github.com/divyang7666/c879940376d8bbd60bec193619594a1c>)

I have generated the data for Counties in UK from the site called generate data.com and other Is github.com from where I got the counties for IOWA. I have added that columns into the UK and IOWA datasets to know the accident-prone area with meeting all the assumptions. Here is the picture below from where I have taken the data..

Took County names from the google search and enter the names in the mock data site.

Country	County
England	London
England	Bedfordshire
England	Buckinghamshire
England	Cambridgeshire
England	Cheshire
England	Cornwall and Isles of Scilly
England	Cumbria
England	Derbyshire
England	Devon
England	Dorset

I have curated these data from above links and coerced them according to the different columns in the data sources like weather conditions, lightning conditions, casual severity, road conditions and different number of injuries and fatalities. In IOWA dataset, some columns were missing and I have added that columns in cleaning process to match and interlink between UK and IOWA accidents table in which all the assumptions which is related to the columns are met even I have deleted some columns from the IOWA dataset. For the third source I have used it to find the sentiment analyses in which I haven't fabricated any data. After the cleaning process like removing blank spaces, remove duplicates, extra rows, column matching, identifying keys to integrate with each other.

At this process, my dataset is ready for modelling so I have extracted these three sources into SSIS (integrating services).

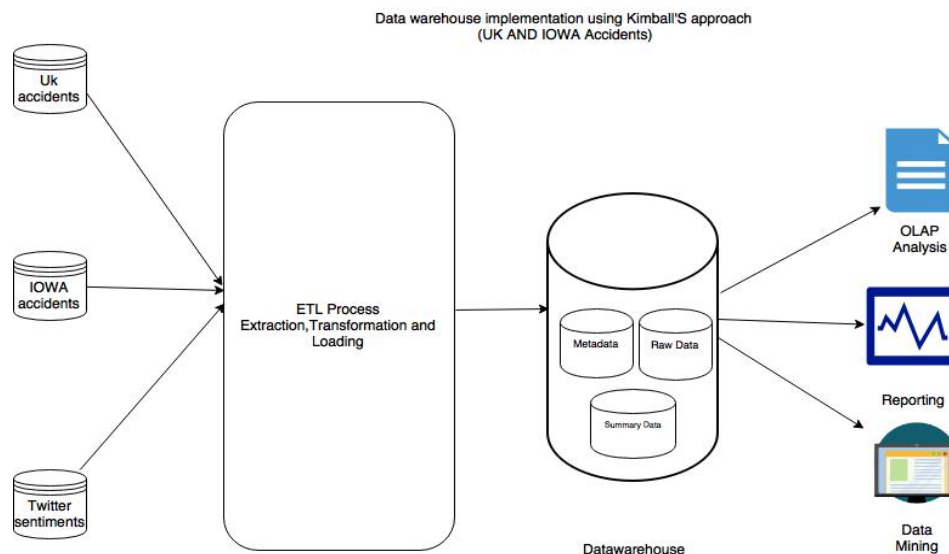
2.Implementation and Architecture

Data warehouse is made from large datasets but before implementing it, it is very crucial to know which approach is appropriated for the organization data. Data warehouse is mainly implemented by two approaches one is developed by Ralph Kimball and other by Bill Inmon or we can use mixed approach. Kimball approach suggest to develop the warehouse in bottom-up approach that means data marts and reports are build first then have to merge the data to make a centralized warehouse. Inmon approach is totally opposite of Kimball approach in which he suggested to implement in top-down approach as first thing is to segregate the data which should be in normalized form to create data marts for the distinct subject areas. It is impossible to tell which approach is correct as both the approaches are beneficial depending upon the needs and granularity of the data (Inmon, W.,2005).

In this project, I have followed the Kimball's approach. First I have extracted the large datasets and merged all the databases into the ETL process by making dimensions (data marts into staging area) and loaded into the fact table which makes the centralized data warehouse required for the business intelligence queries. I have used Kimball approach because of the following reasons:

- 1.This Data warehouse is quick to set up and build which will deliver the first phase quickly where if I have used the Inmon approach it would be more complex and time taking project.
- 2.Business queries work better in any BI tools as star schema can be easily understood by the business users like getting to know the results of accidents areas and injuries.
- 3.Focused on increasing the need of the data warehouse for the government purpose to minimize the accidents.
- 4.Easily to drill down as all BI tools works great to generate a report which is built through dimensions like I have drill down from country of IOWA to the counties of it and the address where accidents happened explained later in this project.
- 5.This approach is process oriented which targets only on small and specific fields like accidents and giving all the information about it.

Here is the diagram which explains the dimensional model of accidents data warehouse (Breslin, Mary.,2004).



Procedure for DWH:

1. Collected data sets from various sources.
2. Designed the data warehouse (DWH)
3. Designed ETL (Extract, Transform and Load) steps
4. Developed ETL
5. Loaded into the warehouse

Software tools I used for this process:

1. SQL Server Management Studio (SSMS)
2. SQL Server Integration Service (SSIS)
3. SQL Server Analysis Service (SSAS)
4. SQL Server Reporting Service (SSRS)
5. Tableau
6. Power BI

3. Data warehouse Data model

I have used star schema to implement the data warehouse of UK and IOWA accidents. It is easier and faster than snowflake schema and remove the complexities to make the multidimensional model. In star schema, there is one fact table which is linked with dimension tables forming a star. There are a lot of benefits of using star schema as follows:

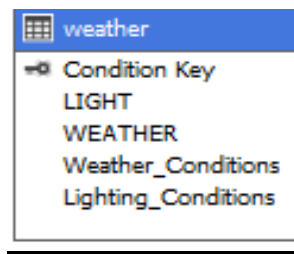
1. Makes the Extract, Transform Load process easier in implementing the data warehouse.
2. Metric Analysis is easier in star schema.
3. Simple query execution.
4. Dimension tables are directly connected to the fact table which increases the data efficiency.

(Kimball, R. and Ross, M., 2011).

Fact contains measures and foreign keys linking with the dimensions which have primary keys that are unique and cannot have null values with attributes like road surface, road class, county, age, sex. Here is the image of the star schema describing the facts and dimensions.

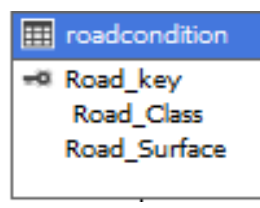
Here the tables which are in blue colour are my dimensions and the yellow table is my fact table which is connected to the dimensions. These dimensions and facts are made from three files which I have taken from three different sources. There is no need for time dimension in my data set as all the accidents were reported in 2015 and there is no matched data and any query to relate the UK and IOWA datasets.

1. Weather is the dimension(dim) table in which it has condition key as primary key which has unique value and rest of them are my attributes, Light and weather first two attributes has numeric values which are available in fact table too and this dimension can be very useful in business queries. In which weather and lightning conditions injuries happened the most.



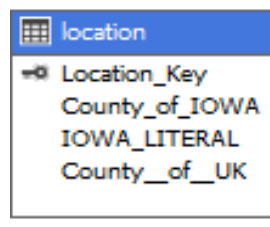
weather	
Condition Key	
LIGHT	
WEATHER	
Weather_Conditions	
Lighting_Conditions	

2.In road condition dimension(dim) road key is my primary key which is connecting to fact table. Road class and surface are attributes which is used in queries to find on which road type maximum number of vehicles are crashing.



roadcondition	
Road_key	
Road_Class	
Road_Surface	

3.Location(dim) has location key as a primary key and three of them are attributes which are used in reporting to identify which area of IOWA has maximum number of vehicles crashed.



location	
Location_Key	
County_of_IOWA	
IOWA_LITERAL	
County_of_UK	

4.Sentiment dim has UK tweet id as a primary key which is connecting to fact table. All the attributes are in this table which are available in the fact as measures to find the sentiment analyses between two different places UK and IOWA. The is one more option which I could have used is that there can be two different sentiment tables one for UK and for IOWA but tweet id is just connecting to the fact table it's not used in business queries so I will be using one table for the queries. It can be depicted as two but to reduce the complexities I used one table.

sentiment
UK_tweet_id
IOWA_tweet_id
Pos_Sent_IOWA
Neg_sent_IOWA
Pos_sent_uk
UK_neg_sent

5.Uk details is other dimension which has collision details regarding injuries, fatalities, vehicles, type of vehicles, age, sex of casualties, causality class, vehicles that destroyed completely and casual severities. Here reference key is the primary key which is connecting to the fact table. Other attributes are used to find the number of fatalities, injuries linked with other tables which are covered in reporting section to find the queries.

ukdetails
Reference_Key
UK_Number_of_Vehicles
UK_Fatalities
Casualty_Class
UK_Injuries
UK_Casualty_Severity
Sex_of_casualty
Age_of_Casualty
UK_vehicles
Type_of_Vehicle

6.IOWA details is last dimension which is formed by IOWA accidents table where crash key is the primary key connecting to the fact table. There are a lot of attributes like gender, number of vehicle, injuries, causality severity, vehicles which were destroyed in accidents and IOWA fatalities which are joined with other tables in reporting section to find the business queries giving the right information to the government.

iowadetails
CRASH_KEY
Gender
IOWA_Number_Of_Vehicle
IOWA_Injuries
IOWA_Casualty_Severity
IOWA_VEHICLES
IOWA_Fatalities

7.Fact table contains all the foreign keys that is connected to the dimension through primary keys like reference key,crash key,location key,condition key,road key and UK tweet id. Fact table contains

measures like the data which have numeric values like UK and IOWA positive, negative sentiments of both the places. Injuries, fatalities, number of vehicles, vehicles destroyed. These values are coming from dimensions and making the fact table updated by applying inner join.

Facttable
Reference_Key
CRASH_KEY
Location_Key
Condition_Key
Road_key
IOWA_Fatalities
IOWA_Injuries
IOWA_Number_Of_Vehicle
IOWA_VEHICLES
Neg_sent_IOWA
Pos_Sent_IOWA
Pos_sent_uk
UK_neg_sent
UK_Fatalities
UK_Injuries
UK_Number_of_Vehicles
UK_vehicles
UK_tweet_id

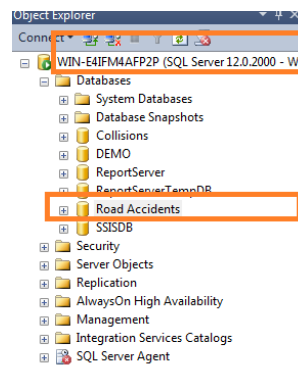
4.ETL strategies

4.1 Extract

4.2 Transform

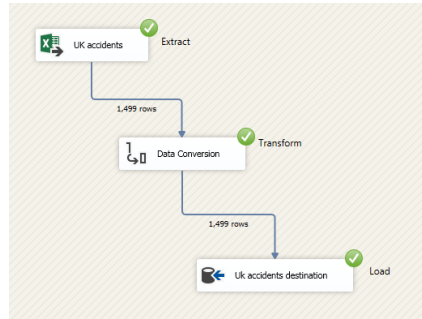
4.3 Load

In IT field ETL is very important process which is used in database system to fetch the data from heterogeneous as well as homogenous data sources. Different databases are needed to incorporate which happens in ETL process (Kimball, R., & Caserta, J.,2004). Extraction of data is an important process which makes the work easier and then transformation (Cleaning) and loading of data can be done in SSIS tool described above rather than using informatica. SSIS can handle multiple different data sources in single package which makes the work easy to manage. After the extraction process it comes transformation where there is **no compulsion of staging platform as its data execution is done parallel**. Cleaning is done in this process like by using data conversion or changing data types and then can load the data into the destination. However, SSMS also used the ETL so first I have created the database in the SQL management studio then connected and implemented ETL (Kimball, R., & Caserta, J.,2004).



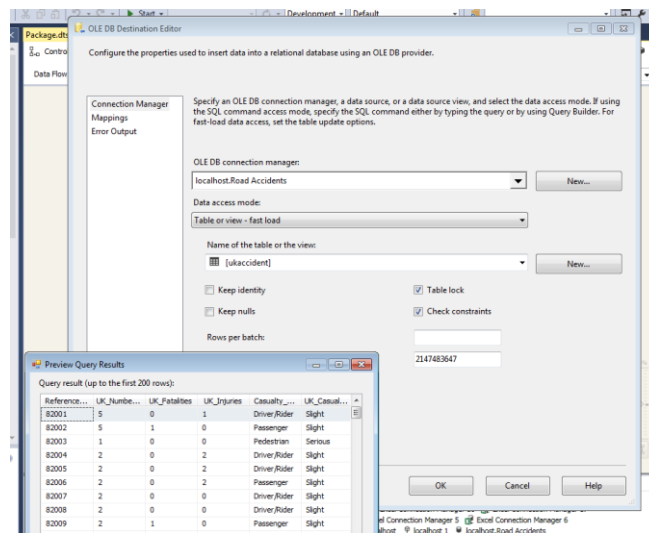
Here **WIN-E4IFM3AFP2P** is the name of the server where all the databases are connected to each other. **Road accidents** is my name of the database.

(I have used script task to pop up the message box and then truncate query in the execute task to truncate the tables whenever it gets started then joined all the staging databases to that)



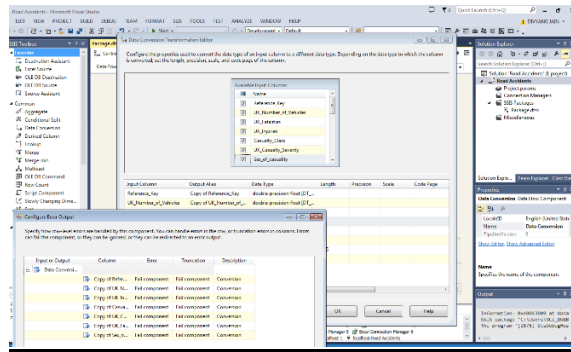
ETL PROCESS

1.Extraction- I have shown the data sources and data sets in the introduction which are cleaned with no spaces and duplicates in the primary keys. Now these excel file will be loaded into the SSIS and will go through the ETL process. There are three excel files so I am showing the one file how I have extracted the data.UK accidents, IOWA accidents and the twitter sentiments of UK and IOWA.



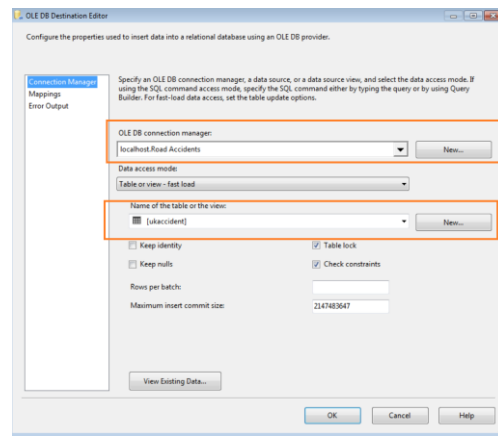
Source Extraction

2.Transform-It is done to clean the data and if there is any need to change the data type or there is any column which is missed we can edit that in this process. Here is the pic where I have changed the data types and cleaned the data for the loading process. We can add derived column and many more too.



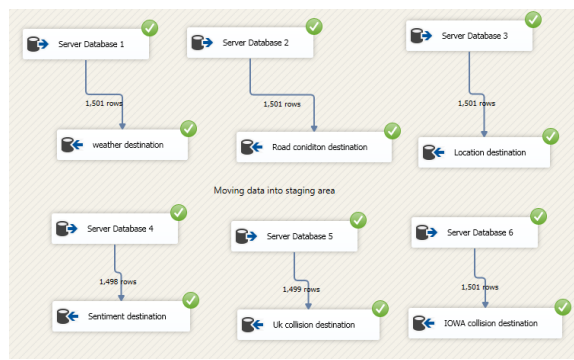
Transforming the data

3.Load-After the data is cleaned and transformed, it needs to be loaded somewhere and I have loaded it into the SSMS by adding Ole DB destination and the table will be added in the road accidents database but the important point here is to see the mappings that all are properly done and now read to load into the table. This is an automation where we don't need to create table in SSMS it will create automatically.

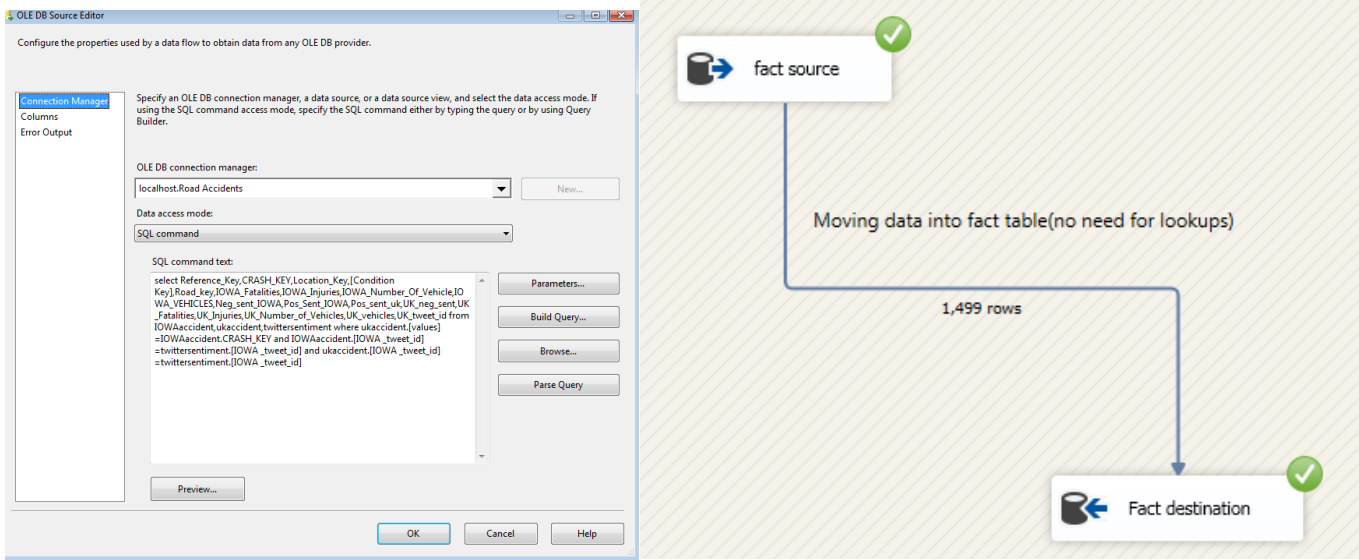


Loading Process

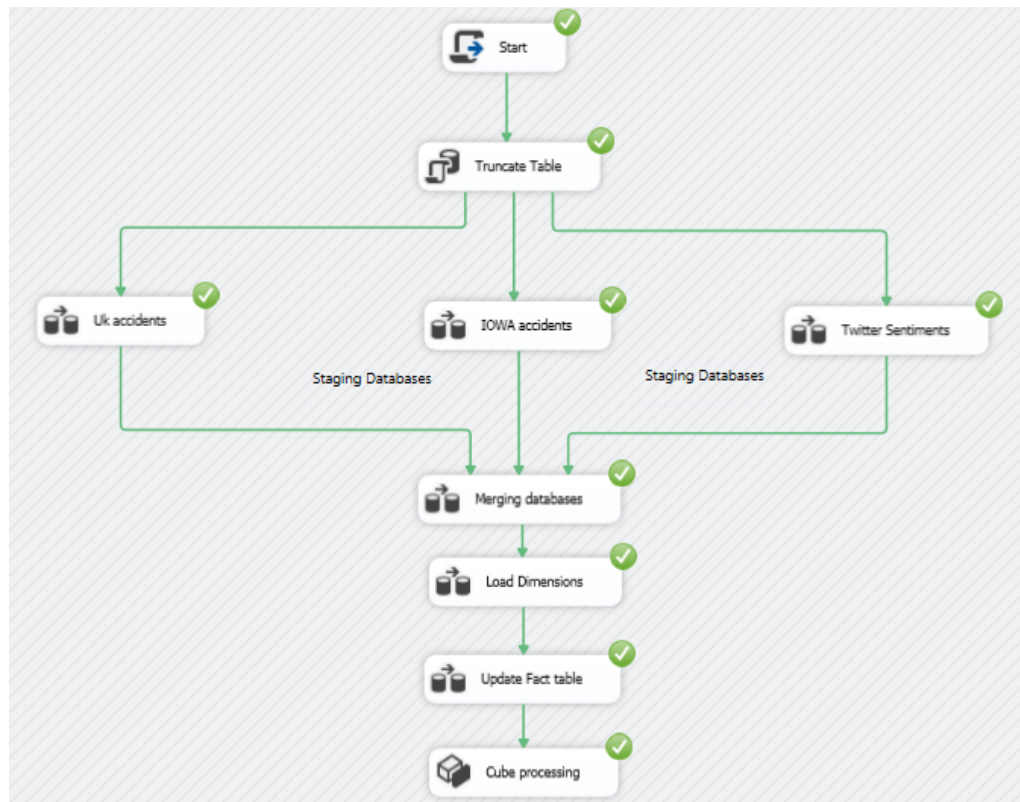
Load Dimensions- After the source tables, I have created the dimension through SQL code by putting in it in oledb source and loaded the dimensions into the oledb destination by automating the tables. Weather, road condition, location, sentiment, UK details and IOWA details are my dimensions.



After updating the dimensions, I have updated the fact table with the **SQL code** and destination of the fact table is SSMS where all the rows and columns will be automated.



Finally, fact table is deployed and there is no need for lookups as all the data are coming from dimensions which is needful through SQL code.



All the values are working and populated into the SSMS and cube as I have used the analytical processing which will be automated to the SSAS directly.

[illegible]

Values are populated and all the tables are created in SSMS and I have highlighted the vales accordingly.

1.Blue colour tables shows the staging database tables which I extracted from excel files.

IOWA accident, UK accident, Twitter Sentiment.

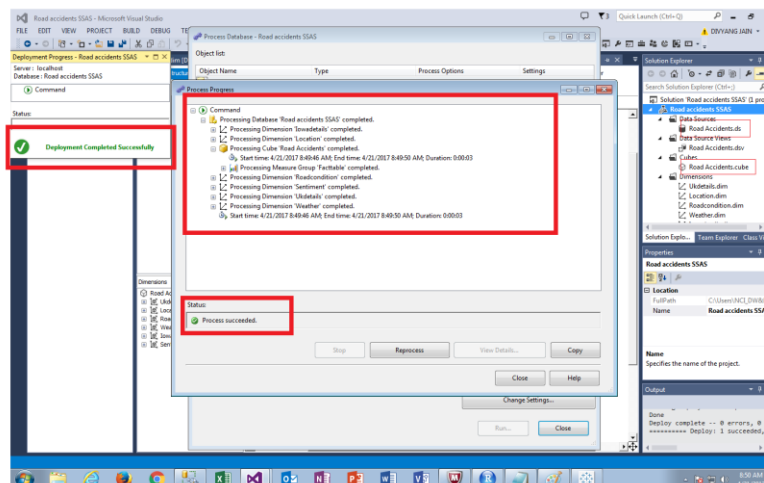
2. Orange colour tables shows the dimensions which have created from the excel files in loading dimensions.

3.Red colour tables shows the fact table which will integrate all the dimensions.

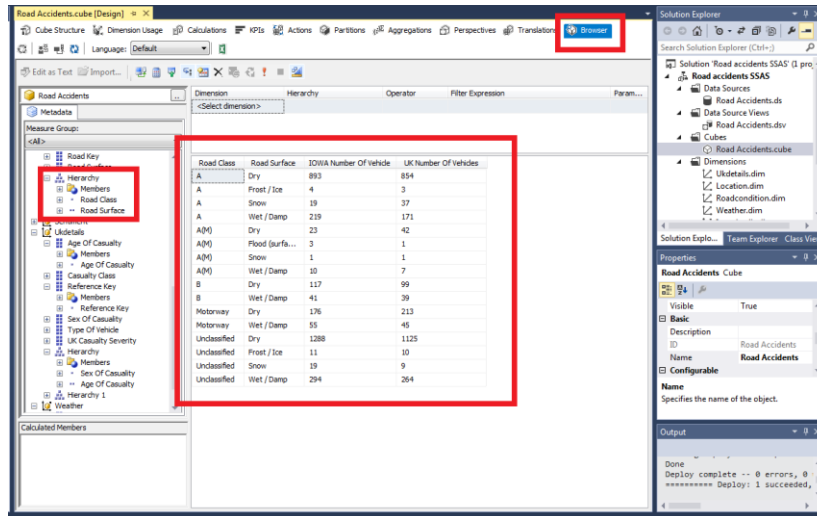
4Yellow table shows the populated values of fact table.

Cube Deployment through SSAS:

In analytical services first I have imported the data source and data source view, through **SSIS cube was automated** and the values of the cube were properly deployed hereafter creating that I have made the **hierarchies** like: 1-sex of causality, age.2. Causality class and causality severity.3 County of IOWA and literal of it 4. Road class and surface. Here before deploying the cube I have made primary keys in the dimensions and made the relation to fact table foreign keys. Now I have process the cube and then deployed.



I have deployed the cube this means there is no error and all the values are shifted to the fact table.



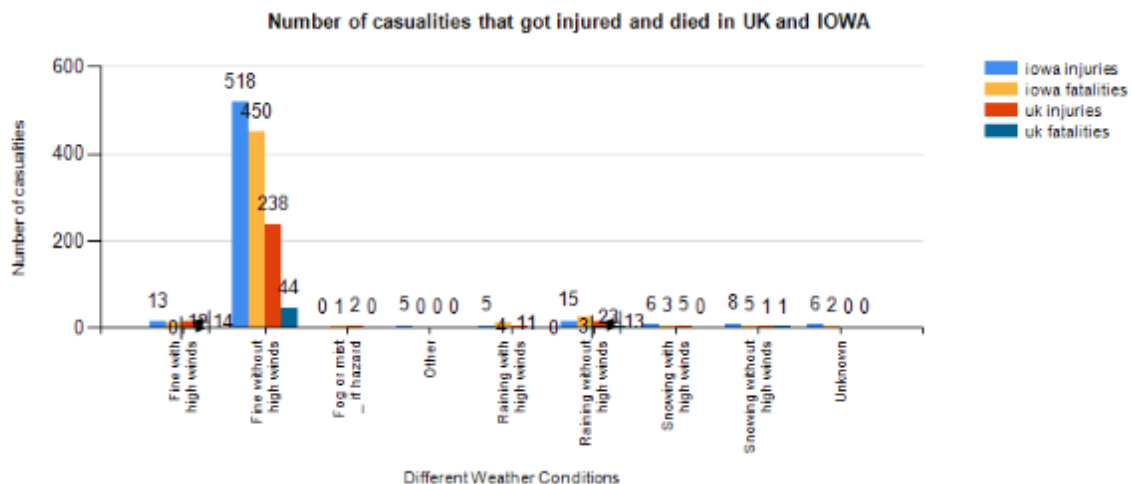
5.Applications of Data Warehouse (Reporting)

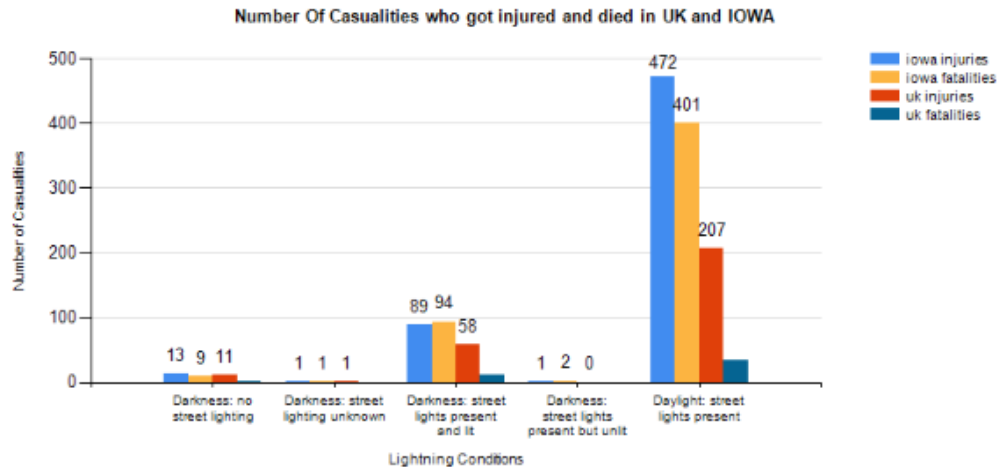
Reporting plays a key role as it is the vital step which includes the analyses of business queries and that is the best application of data warehouse. Reports can be made after deploying the cube as the cube has fact table and dimension table values which will integrate the values and make data visualization. There are BI tools which I used to find the queries on the data warehouse and that are: **1: SSRS 2: Tableau and Power BI.**

Research Queries Case study: -

1.SSRS Report:

1.In which weather and lightning conditions, the maximum number of injuries and fatalities occurred in UK and IOWA?





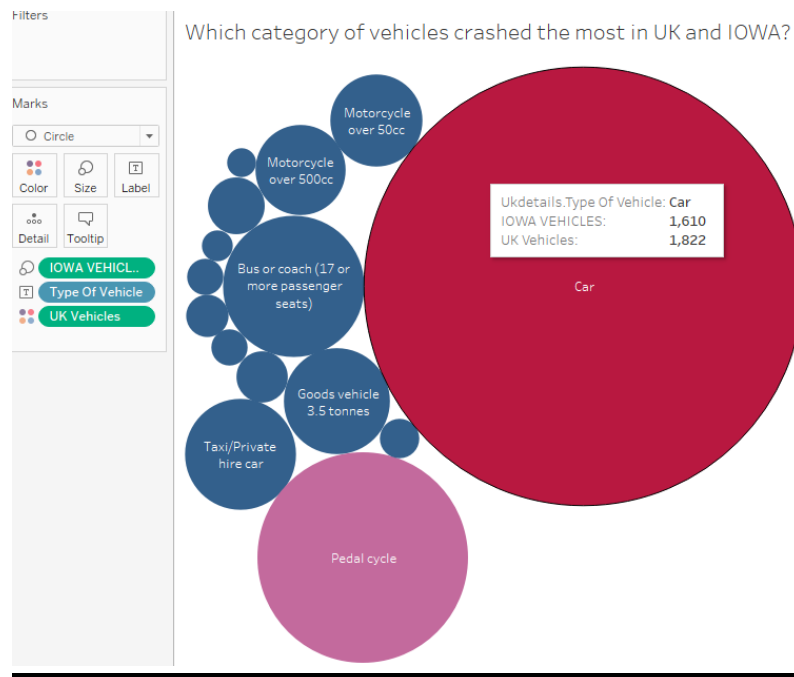
The first graph shows that injuries, fatalities happened mostly in fine without high winds weather and in rest of the weather conditions it's not too much. The second graph shows the number of injuries, fatalities happened in different lightning conditions and in daylight street lights present it happened the most. Both the graphs indicate that IOWA injuries and fatalities are more than UK injuries and fatalities.

Data sources –UK details, IOWA details and weather dimensions used linking with fact table.

2.Tableau:

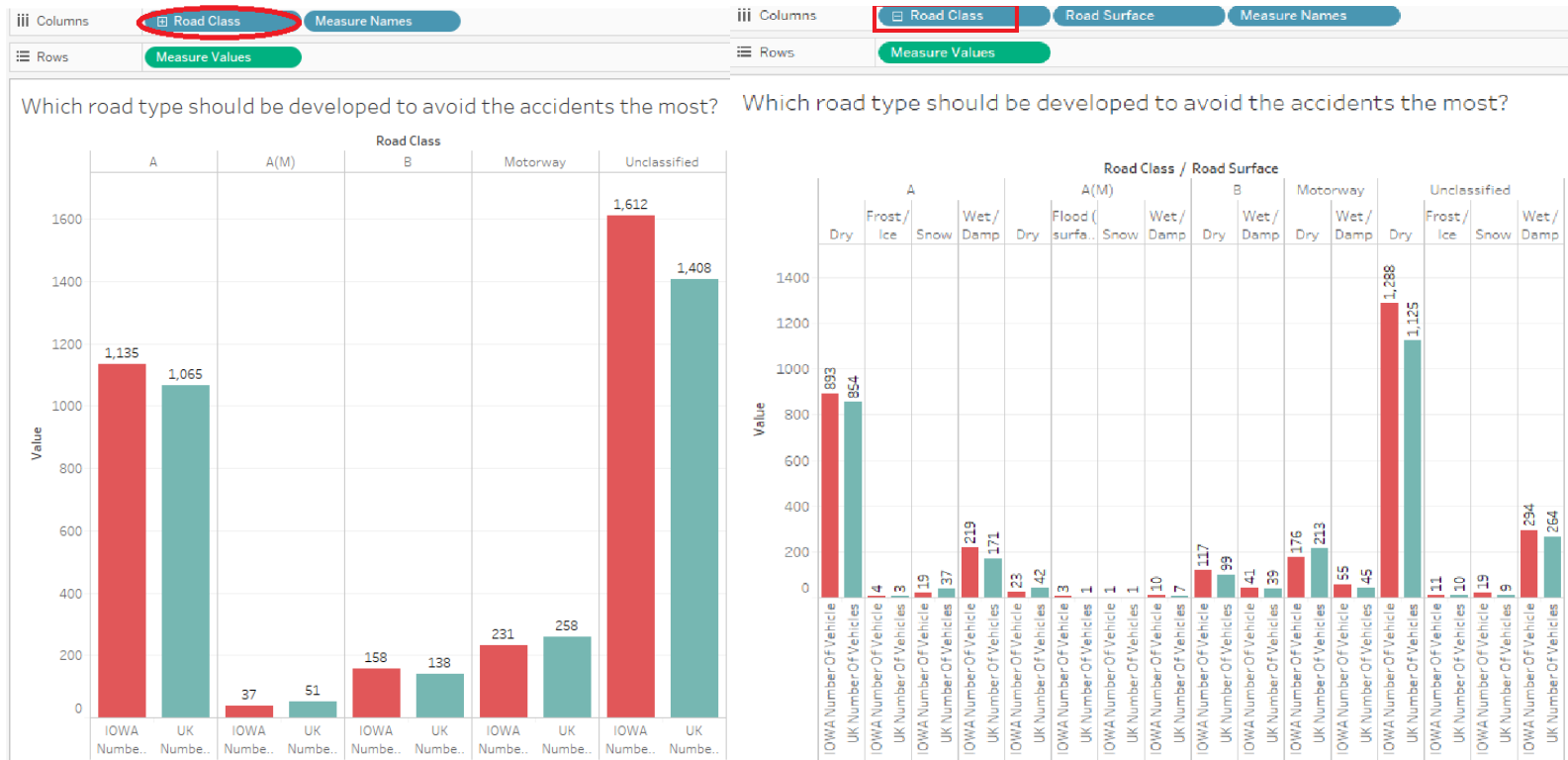
I have imported the cube first into the tableau from SSAS by connecting through the local host, it automatically showed the measures and dimensions on the left panel of the worksheet. I have used it because it shows a lot of flexible queries and good range of visualization then others.

2.Which category of vehicles crashed the most in UK and IOWA?



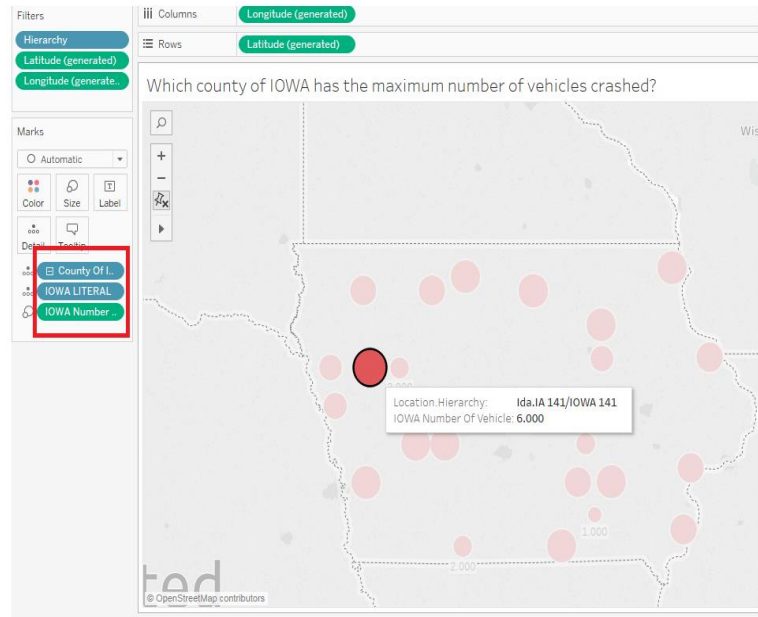
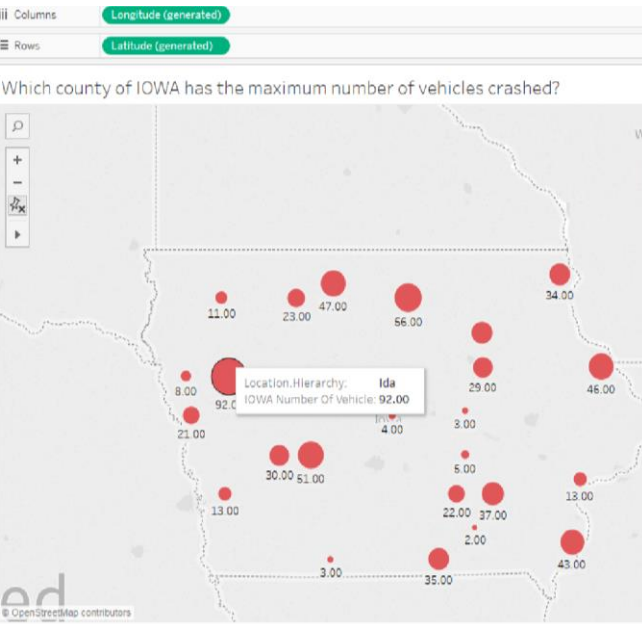
As from the figure we can see **car** shows the maximum vehicles destroyed completely in both the areas of UK and IOWA. Among them UK has more than IOWA vehicles that destroyed in accidents that is around 1610 in IOWA and 1822 in Kythera are a lot of categories which shows slightly different from car that is pedal cycle, bus or coach, motorcycle and many types of categories in the graph. Government must look forward into this information and take some measures to avoid the accidents. Data Sources-UK details and IOWA details dimension linking with fact table.

3. Which road type should be developed to avoid the number of accidents the most?



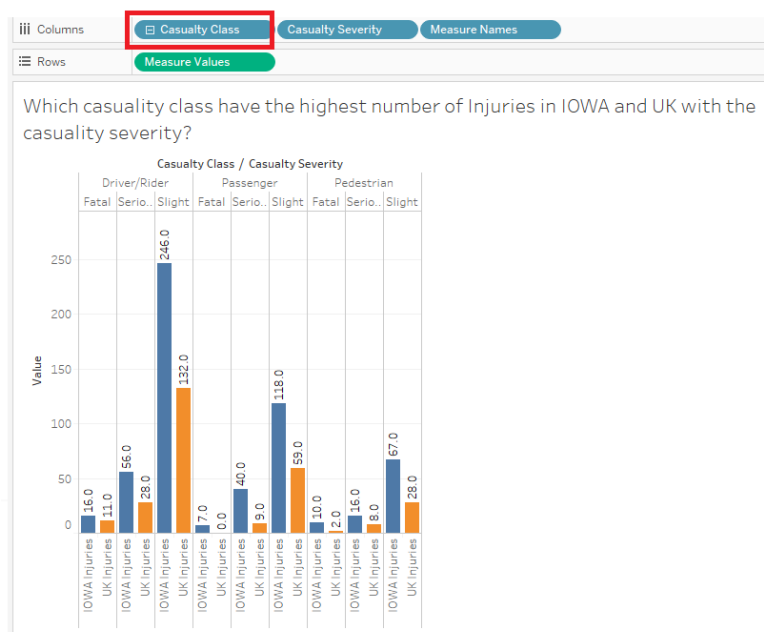
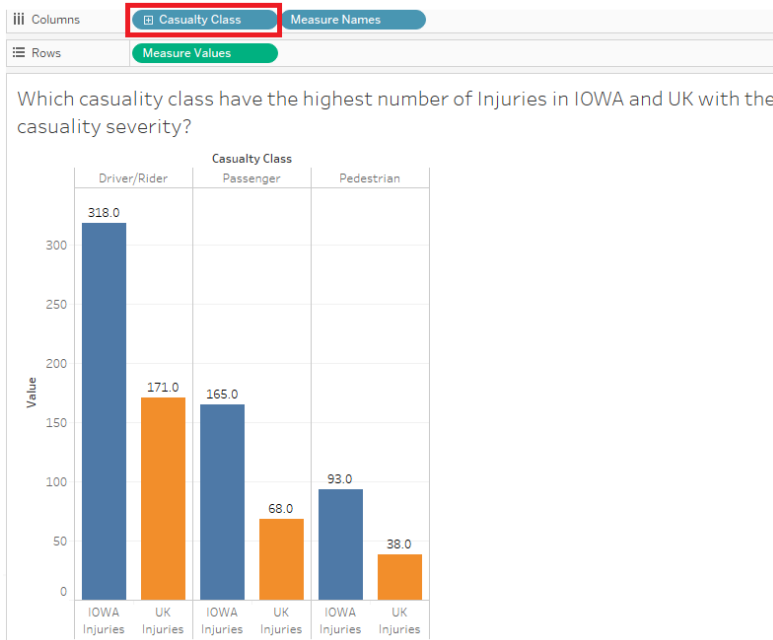
These two graphs show the road condition with the number of vehicles collided in UK and IOWA. In both the places vehicles crashed mostly on unclassified road which is the one of the type of road class and in IOWA has the maximum number of vehicles crashed then UK (1612, 1408) respectively and in A(M) type road it has the minimum number of road collisions. To understand it better I, have **drill-down** The information to road surface which shows in which road class having what kind of road surface have the most number of accidents, it shows the dry surface of unclassified road class has the maximum number of vehicles crashed and the A(M) category have lowest. So, government should build the A(M) roads to avoid the number of accidents happened in both the places. Data source-Road dimension and IOWA, UK details linked with fact table.

4. Which county of IOWA has the maximum number of vehicles crashed?



The graph shows the different counties of IOWA which is one of the biggest state of US having the most number of crashed vehicles (92) in IDA county of IOWA. I have **drill-down** to the addresses of it where it happened, each address shows the different number as the total it says is 92 and one of the address has 1A 141/IOWA 141 6 crashed vehicle. Government should declare this area as an **accident-prone area** and they should start a lot of measures which prevents the accidents in this area. Data source-Location county, IOWA details county linked with fact table.

5. Which causality class have the highest number of injuries in IOWA and UK with the casual severity?



In most number of accidents, driver/rider got injured the most in both the areas of UK and IOWA. Pedestrian has the lowest number of injuries and if I **drill-down** to the casual severity of the causality class it shows that maximum accidents was in which driver/rider got injured is not so serious as it shows the slight category in the causality severity. So very less number of driver got the fatal severity and serious condition so there were very less chances to get killed in the accidents. Car companies should develop more safety measures for the person sitting in the car.

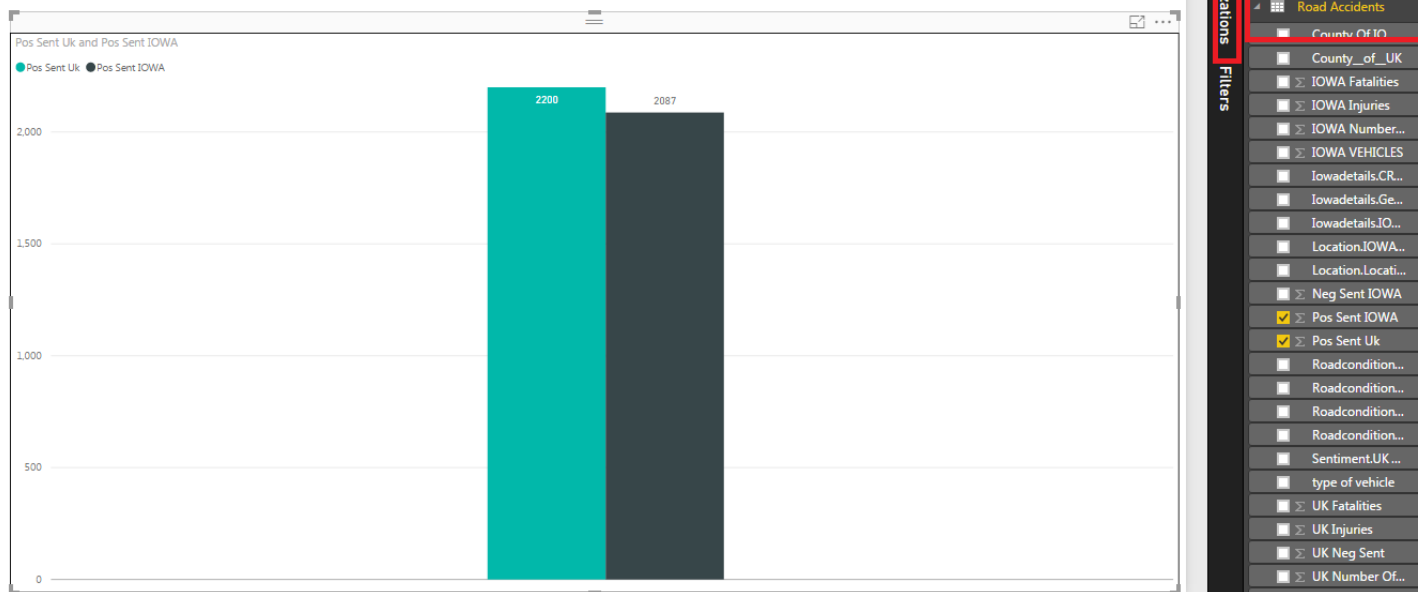
Data source-IOWA and UK details which has causality severity with the causality class linked with fact table.

3.POWER BI

Directly imported my SSAS cube to power BI tool by going into the get data option and linking with the local host then clicked on Road accidents SSAS cube and showed the queries. It is simple to show the queries here then other BI tools.

6.Which place is safer in terms of accidents in public opinion?

Twitter Sentiment analyses on UK and IOWA Accidents



The graph shows the **twitter sentiment analyses** between two places. It shows that UK got the more positive response then IOWA stating that UK people is happier and satisfied then IOWA as there may be less number of accidents.UK got 2200 positive tweets and IOWA got 2087 votes stating that people from UK are more satisfied then IOWA.

Conclusion-Government must initiate some actions to make IOWA accident free area.

Data Sources-Sentiment dimension with the fact table.

References

- (1) Breslin, Mary.,2004. “Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models” Business Intelligence Journal, Winter 2004. Accessed May 22, 2016.
- (2). Inmon, W.,2005. *Building the data warehouse*. 1st ed. Indianapolis, IN: Wiley Pub.
- (3). Kimball, R., & Caserta, J.,2004. *The data warehouse ETL toolkit*. John Wiley & Sons.
- (4). Kimball, R. and Ross, M., 2011. *The data warehouse toolkit: the complete guide to dimensional modelling*. John Wiley & Sons.
- (5). Matloff, N., 2011. *The art of R programming: A tour of statistical software design*. No starch press.