

Data Warehousing and Business Intelligence Project

on

Natural Gas and Oil Consumption, Production and Trade based
on Reserves, Population and Co2 Emission

Pavan Kumar Sudhakar
x17126738

MSc/PGDip Data Analytics – 2018/9

Submitted to: Dr. Simon Caton

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing



Student Name:	Pavan Kumar Sudhakar
Student ID:	x17126738
Programme:	MSc Data Analytics
Year:	2018/9
Module:	Data Warehousing and Business Intelligence
Lecturer:	Dr. Simon Caton
Submission Due Date:	26/11/2018
Project Title:	Natural Gas and Oil Consumption, Production and Trade based on Reserves, Population and Co2 Emission

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	November 26, 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used L^AT_EX template
- ☒ Three Business Requirements listed in introduction
- ☒ At least one structured data source
- ☒ At least one unstructured data source
- ☒ At least three sources of data
- ☒ Described all sources of data
- ☒ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☒ Inserted and discussed star schema
- ☒ Completed logical data map
- ☒ Discussed the high level ETL strategy
- ☒ Provided 3 BI queries
- ☒ Detailed the sources of data used in each query
- ☒ Discussed the implications of results in each query
- ☒ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Natural Gas and Oil Consumption, Production and Trade based on Reserves, Population and Co2 Emission

Pavan Kumar Sudhakar
x17126738

November 26, 2018

Abstract

Natural Gas and Oil are prominent energy sources across the world in the production of electricity and in domestic consumption. The usage of these energy sources has increased each year, and hence the production is increasing in order to meet the growing demand.

The production of energy sources depends on the resources available, and the consumption is correlated directly with the population. The usage of this energy sources has a major setback because of its property of emitting Co2. The emission of Co2 by fossil fuels is one of the major concerns around the world as this is a prime contributor for many diseases particularly lung cancer and Asthma.

This Data warehousing project is made to research the relationship between Natural Gas and Oil production, consumption, resources available, population and the Co2 emitted by each energy sources around the world. This project also particularly analyses the total amount of Co2 emitted by Oil and by Natural Gas individually, so that the benefits of using one source of energy over the other can be compared. The future work of this research will extend to compare the use of more fossil fuels and their pros and cons.

1 Introduction

The use of fossil fuels for domestic consumption and in the production of electricity is a common practice across the world. There are quite a few fossil fuel types like Petroleum, Natural gas, coal etc. Out of these Oil and Natural gas are the most commonly used fuel types. But there is an important factor that needs to be considered while using these fuels as a source of energy and that is Carbon dioxide (Co2) emission. The emission of Co2 is a serious problem throughout the world since it contaminates the environment in the best possible way. The major problems are Global warming which is causing the Atlantic ice to break and dissolve which is increasing the ocean level each day, already several small islands sunk in the ocean. There are also many health-related problems like Lung Cancer and Asthma for which the emission of Co2 is a major cause. In this Data warehousing project suitable datasets for Production and consumption of Oil and Natural Gas by each country, Oil and Gas reserves available across world and amount of

Co2 emitted in mtCo2 by Oil and Gas by each country has been obtained for 2012 to 2016 period and formatted in order to facilitate a Data warehousing model.

- (Req-1) Comparing the consumption of Natural gas and Oil globally and the emission of Co2 by each fuel source year wise and check for a correlation.
- (Req-2) Comparing the Production of Natural gas and Oil worldwide based on resources available globally and check for a correlation.
- (Req-3) Comparing the Trade of Natural gas, Oil and Population worldwide and its impact on GDP Growth and check for a correlation among them.

2 Data Sources

5 different data sources have been used to build a data warehousing model in order to analyze the business requirements. Out of which 4 sources are structured and 1 source is Unstructured. The detail explanation of each sources were elaborated in below subsection.

Source	Type	Brief Summary
Enerdata	Structured	Data of Natural gas and Oil production, Consumption by each country from 2012 to 2016
Global Carbon Atlas	Structured	Data of Co2 emitted by each country from 2012 to 2016.
Statista	Structured	Data of Natural Gas reserves worldwide from 2012 to 2016.
Statista	Structured	Data of Oil reserves worldwide from 2012 to 2016.
Kaggle	Structured	Population data by country wise from 2012 to 2016.
World Trade Organization	Unstructured (Image)	Data of region wise GDP growth from 2012 to 2016

Table 2: Summary of sources of data used in the project

2.1 Source 1: Enerdata

The Natural Gas and Oil Production, Consumption and Trade for 44 countries data-set was downloaded from: <https://yearbook.enerdata.net/natural-gas/world-natural-gas-production-statistics.html>. This data-set provides 58 rows and 31 columns of information regarding Gas and Oil production, consumption and Trade for 44 countries.



Figure 1: Publication date for Source 1

This data-set is structured and transformed into 220 rows and 10 columns in order to facilitate into the Data warehouse model and provide options for drill down. This data-set was used in addressing two business requirement by providing information regarding Natural Gas consumption and production. The publication date of this data source is shown in the Figure:1

2.2 Source 2: Global Carbon Atlas

The data-set on Co2 emission by Oil and Gas was downloaded from: <http://www.globalcarbonatlas.org/en/CO2-emissions>. This data-set provides 220 columns and 8 rows of information on Co2 emission by Oil and Gas for 44 different countries and for the period from 2012 to 2016.

Re: Dataset publication date



Anna Peregon <anna.peregon@lsce.ipsl.fr>

16/11/2018 13:24



To: pavankumar s

Dear Pavan Kumar,

Thanks for your interest in using C-Atlas datasets in your project.

Is your question about (national) EMISSIONS data? National emissions published in the Atlas refer to time period 1960-2016. This dataset was published in November, 2017 (during the COP23).

Figure 2: Publication date for Source 2

This data-set was transformed into 220 rows and 9 columns in order to facilitate into the Data warehouse model and provide options for drill down.

This data-set is used to address one business query by providing information on Co2 emission by Oil and Gas. This data can also be used to compare Population, Production with Co2 emission which could be related to each other and a linear model can be drawn and studied. This will be a part of the future work of this model. The publication date of this Source is shown in Figure:2

2.3 Source 3: Statista

The data-set on Natural Gas and Oil reserve was downloaded from: <https://www.statista.com/statistics/273584/distribution-of-natural-gas-reserves-by-region/>
<https://www.statista.com/statistics/271614/oil-reserves-by-region/>

Distribution of proved natural gas reserves worldwide from 1992 to 2017, by region

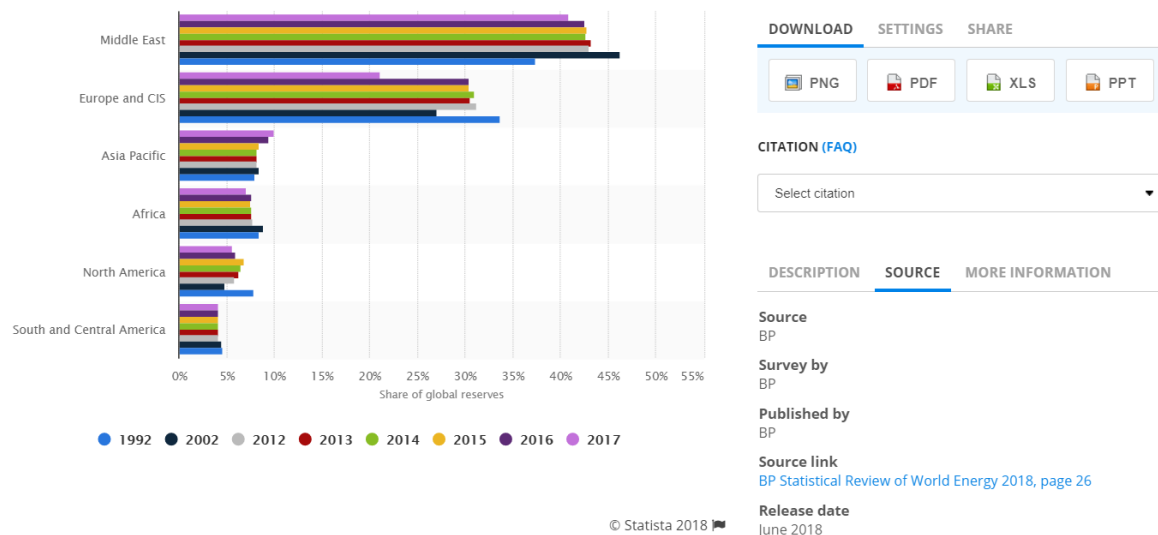


Figure 3: Publication date for Source 3A

Proved oil reserves worldwide from 1992 to 2017, by region (in billion barrels)

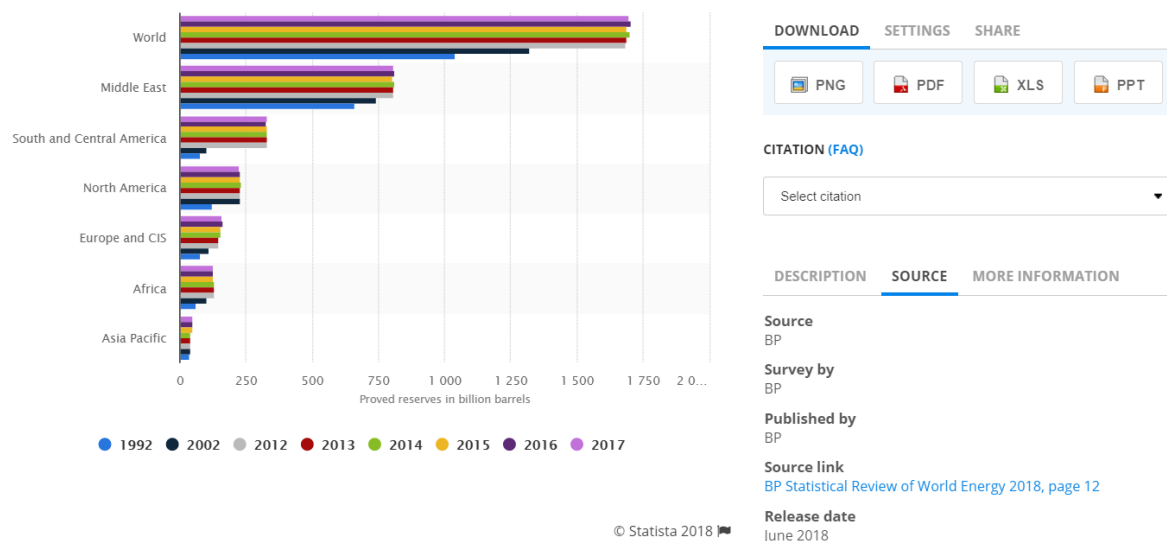
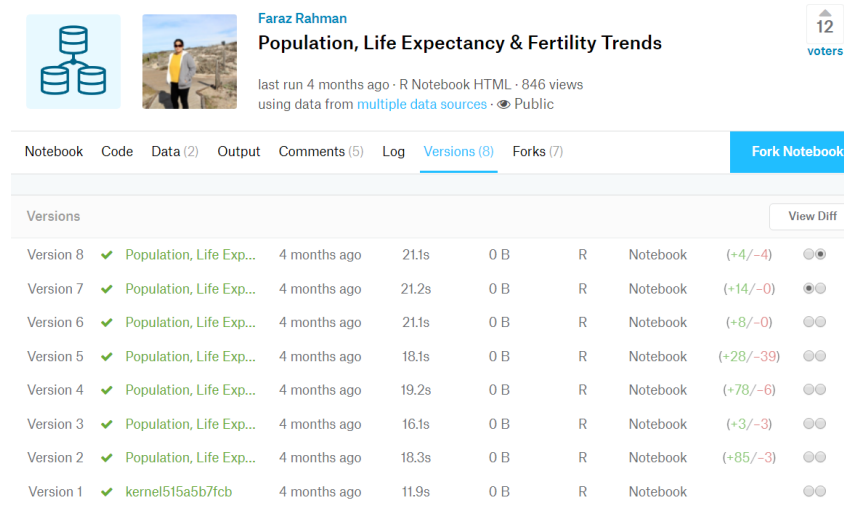


Figure 4: Publication date for Source 3B

This data set provides data about the Natural gas and Oil reserves worldwide in percentage. This data-set provides 7 rows and 9 columns, this was transformed into 50 rows and 6 columns to fit into warehouse model and for Drill down. This source is used in addressing one of three business requirements and the publication date of this source is shown in Figure:3A and 3B.

2.4 Source 4: Kaggle

The data-set on Population human population worldwide was downloaded from: <https://www.kaggle.com/farazrahman/population-life-expectancy-fertility-trends/data> This data-set provides 62 columns and 261 rows of information on world population by country wise and year wise. This source of information is used to address the 3rd business requirement in comparing Fuel Trade with GDP and Population. The published date for this source is given the figure:4



Version	Status	Title	Time	Size	Language	Environment	Changes	Actions
Version 8	✓	Population, Life Exp...	4 months ago	21.1s	0 B	R	Notebook (+4/-4)	View Diff
Version 7	✓	Population, Life Exp...	4 months ago	21.2s	0 B	R	Notebook (+14/-0)	View Diff
Version 6	✓	Population, Life Exp...	4 months ago	21.1s	0 B	R	Notebook (+8/-0)	View Diff
Version 5	✓	Population, Life Exp...	4 months ago	18.1s	0 B	R	Notebook (+28/-39)	View Diff
Version 4	✓	Population, Life Exp...	4 months ago	19.2s	0 B	R	Notebook (+78/-6)	View Diff
Version 3	✓	Population, Life Exp...	4 months ago	16.1s	0 B	R	Notebook (+3/-3)	View Diff
Version 2	✓	Population, Life Exp...	4 months ago	18.3s	0 B	R	Notebook (+85/-3)	View Diff
Version 1	✓	kernel515a5b7fcb	4 months ago	11.9s	0 B	R	Notebook	View Diff

Figure 5: Publication Date for Source 4

2.5 Source 5: World Trade Organization

The data-set on GDP growth worldwide was downloaded from https://www.wto.org/images/img_press/768tbl1_e.png. This data-set provides information about GDP growth worldwide by region wise. This is an unstructured data source which is scrapped from image and transformed into 4 columns and 25 rows to fit into Data warehouse model. This source of information is taken from the web directly every time the SSIS package is invoked to fetch Raw data.

3 Related Work

In Nejat et al. (2015) the author compares the energy consumption, CO2 emissions, and policy in the residential sector for ten countries China, the US, India, Russia, Canada, Iran, South Korea, Japan, Germany, and the UK. Which is accounted for two-thirds of the global Co2 Emission, which is accounting for two-thirds of global Co2 emissions. This paper explains that global consumption has increased by 14 percent from 2000 to 2011. This Paper mainly shows that Co2 emissions were mostly emitted by the Developing countries, this paper also gives an exception for tow developed countries which are America and Japan which showed a steady increase in the Co2 emissions. This paper also involves discussing various energy policies such as energy codes, energy labels, and Incentives that are employed by various countries. Data sets on Co2 emission by ten countries and consumption of energy are taken to make the comparison and the results are achieved.

In Bildirici (2017) author compares the Oil production, consumption and its impact on GDP growth on MENAP countries. This paper uses ANOVA and ARDL methods to nexus the relationship. As per the results of this paper, the economic growth in middle east countries has a positive impact in relation with the Oil consumption. From the results of this paper, it is evident that the GDP growth on Middle East and Africa has a linear relationship with the consumption of Oil. Data sets on GDP growth and Oil consumption in Middle East and Africa has been taken to compare the relationships.

In Dong et al. (2017) author establishes the relationship between Co2 by Natural gas and Oil emission and GDP growth in BRICS countries (Brazil, Russia, India, China, South Africa). Panel unit root, causality, and cointegration tests are run to run this model. This paper predicts that the increase in consumption of natural gas and renewable energy decreases the Co2 emission, which is predicted by the model as For 1 percent increase in Natural gas consumption the Co2 emitted by BRICS countries will decrease by 0.165 percent and 0. 26 percent. Thus, this article supports the usage of Natural gas and renewable source of energy and highlights policy implications in tackling the Co2 emission as well as promoting the usage of Natural gas and Renewable sources of Energy. Dataset used in this paper is Co2 emission by BRICS countries and Consumption of Natural gas and Renewable source of energy in BRICS countries.

In Dong et al. (2018) the author uses Cross- sectional dependence and Slope homogeneity tests to establish the relationship between Emission, Population and Economy. The test results show that population and economy growth at region and country level positively influence the Co2 emission levels. Even though increased use of renewable energy decreases the Co2 emission for all six regions, but in South and Central America, Europe and Eurasia the emission is relatively higher, which is mainly due to low proportion of renewable energy in the primary sources.

In Acheampong (2018) author analyses the nexus between carbon emission, economic growth and energy consumption by applying Panel vector autoregression combined with System-GMM (System generated method of moment). The key observations made with the help of multivariate models are, 1. The Economic growth by global and regional level does not impact the energy consumption. 2. Economic growth has no casual effect on carbon emission with an exception to global and Caribbean-Latin America. 3. The Carbon emission will have a positive effect on economic growth. 4. Energy consumption has a positive impact on economic growth in Saharan-Africa and has negative impact on economic growth at the Global, Middle East and North America (MENA), Caribbean Latin America and Asia and Pacific. Energy consumption has positive impact on carbon emission in MENA and negative impact on carbon emission in Saharan-Africa and Caribbean Latin America. 6. The carbon emission is not dependent on energy consumption with an exception to Global and MENA.

4 Data Model

In this data warehousing model, Kimball star schema approach has been used which has the fact table correlated to the Dimension table with Foreign and primary key relation. In general star schema easy to understand and most of the BI query fetch the required data

easily with the star schema. Compared to other schema star perform well in dataware house modelling. Kimball et al. (2013)

This Data Warehousing project has been created by using Star schema approach. Totally there are three dimensions which are listed below.

1. Dim-Time
2. Dim-Location
3. Dim-Fuel-Type

The Dim-Time dimension has been created to group the data by year wise, this dimension has only one level of the hierarchy and that is Year. The data to build this dimension is taken from all the 5 sources of data, as all the source of data has been transformed in such a way that it has years ranging from 2012-2016.

The Dim-Location dimension is an important dimension among all which groups the data by region and country wise, it has two levels of hierarchy. The region values are taken from Energy, Co2 emission, Population data sources as it has data for region and country wise, and Reserves and GDP-Growth sources have data only by Region wise. The Fuel-Type dimension has one level of the hierarchy and that is Fuel Type (Gas/Oil). Data for this dimension are taken from Enerdata, Co2-Emission and Reserves sources.

A maximum of 4 levels of drill down option are made available and that shall be used wisely to query the warehouse model. Year–Fuel–Type–Region–Country, this level is possible while using Energy and Co2 data sources only.

The number of levels depends on the type of source we use to query, the number of level decreases to 3 if we are using Reserves data source, as this source doesnt have data by country wise. In case of using GDP-Growth data source, the drill down options available are: Year– Region. For population the levels are: Year–Region–Country.

For example, let us take the 1st business requirement which is Comparing the consumption of Natural gas and Oil globally and the emission of Co2 by each fuel source year wise and check for a correlation. Here the consumption of Fuels is compared against the Co2 Emission by respective fuels. The fact measure Consumption and Co2 Emission has a full level of drill down that can be used. Both variables have year wise data so Dim-Time can be used, and Region and Country wise data so Dim-Location can be used, and they both have values for Gas and Oil and hence Dim-Fuel-Type can also be used. This makes this comparison very effective and precise.

But if we take the measures Production and Reserves, Dim-Fuel-Type, Dim-Time can be used fully but Dim-Location can be used with only one level of hierarchy because Reserves measure does not have values by country wise. The Star-Schema which is achieved using these dimensions can be seen in figure:5

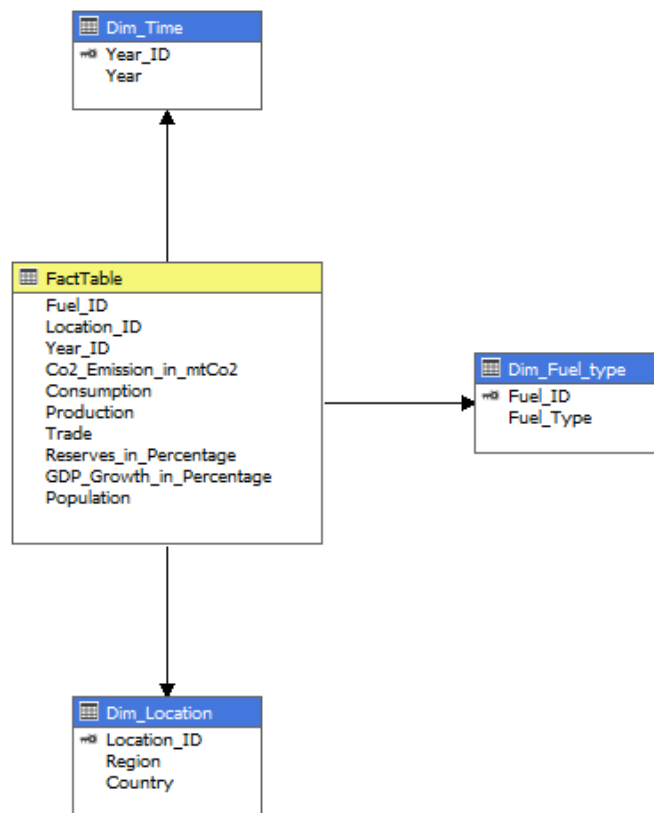


Figure 6: Energy - Star Schema

5 Logical Data Map

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 9

Source	Column	Destination	Column	Type	Transformation
1	Year	DimTime	Year	Dimension	Year data was transformed from column names into rows using gather function so that it can be used in Dimensions and drill downs. Data was filtered only for the period 2012 to 2016.
1	FuelType	DimFuelType	FuelType	Dimension	Data from 2 different files has been merged to produce this dimension value to facilitate drill down.
1	Fuel-ID	DimFuelType	Fuel-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on FuelType column to create unique ID for Fuel types
1	Region	DimLocation	Region	Dimension	This column was made from R code by grouping the available countries taken from the source to facilitate higher level of drill down in 'Location' Dimension.
1	Region-ID	DimLocation	Region-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on Region column to create unique ID for available regions.
1	Country	DimLocation	Country	Dimension	44 countries have been chosen from the source and transformed to accommodate year wise which resulted to 220 rows.
1	Country-ID	DimLocation	Country-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on Country column to create unique ID for available regions.

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
1	Consumption	FactTable	Consumption	Fact	Consumption data of Natural gas and Oil related to chosen 44 countries were selected from 2 different files and transformed to fit into respective countries by year wise.
1	Production	FactTable	Production	Fact	Production data of Natural gas and Oil related to chosen 44 countries were selected from 2 different files and transformed to fit into respective countries by year wise.
1	Trade	FactTable	Trade	FactTable	Trade data of Natural gas and Oil related to chosen 44 countries were selected from 2 different files and transformed to fit into respective countries by year wise.
2	Year	DimTime	Year	Dimension	Year data was transformed from column names into rows using gather function so that it can be used in Dimensions and drill downs. Data was filtered only for the period 2012 to 2016.
2	FuelType	DimFuelType	FuelType	Dimension	Data from 2 different files has been merged to produce this dimension value to facilitate drill down.
2	Fuel-ID	DimFuelType	Fuel-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on FuelType column to create unique ID for Fuel types
2	Region	DimLocation	Region	Dimension	This column was made from R code by grouping the available countries taken from the source to facilitate higher level of drill down in 'Location' Dimension.
2	Region-ID	DimLocation	Region-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on Region column to create unique ID for available regions.
2	Country	DimLocation	Country	Dimenion	44 countries have been chosen from the source and transformed to accommodate year wise which resulted to 220 rows.

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
2	Country-ID	DimLocation	Country-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on Country column to create unique ID for available regions.
2	Co2-Emission-in-mtCo2	FactTable	Co2-Emission-in-mtCo2	FactTable	Co2 emission data of Natural gas and Oil related to chosen 44 countries were selected from 2 different files and transformed to fit into respective countries by year wise.
3	Region	DimLocation	Region	FactTable	6 regions were obtained from Statista data source which was then transformed into 5 regions by combining Africa and Middle East as one entity.
3	Year	DimTime	Year	Dimension	Year data was transformed from column names into rows using gather function so that it can be used in Dimensions and drill downs.
3	FuelType	DimFuelType	FuelType	Dimension	Data from 2 different files has been merged to produce this dimension value to facilitate drill down.
3	Fuel-ID	DimFuelType	Fuel-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on FuelType column to create unique ID for Fuel types
3	Region-ID	DimLocation	Region-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on Region column to create unique ID for available regions.
3	Reserves-in-percentage	FactTable	Reserves-in-percentage	FactTable	Natural gas and Oil reserves data related to chosen 5 region were selected from 2 different Statista files and transformed to fit into respective countries by year wise.
4	Year	DimTime	Year	Dimension	Year data was transformed from column names into rows using gather function so that it can be used in Dimensions and drill downs.

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
4	Region-ID	DimLocation	Region-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on Region column to create unique ID for available regions.
4	Region	DimLocation	Region	FactTable	6 regions were obtained from Statista data source which was then transformed into 5 regions by combining Africa and Middle East as one entity.
4	Country-ID	DimLocation	Country-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on Country column to create unique ID for available regions.
4	Country	DimLocation	Country	Dimenion	44 countries have been chosen from the source and transformed to accomodate year wise which resulted to 220 rows.
4	Population	FactTable	Population	FactTable	Population data obtained from the source file were filtered to get population values for only chosen 44 countries and transformed to fit into each year accordingly.
5	Year	DimTime	Year	Dimension	Year data was transformed from column names into rows using gather function so that it can be used in Dimensions and drill downs.
5	Region	DimLocation	Region	Dimension	Each of the 5 regions were scrapped by using <code>substr()</code> function by using position of the region names from the whole text obtained from the Source Image.
5	Region-ID	DimLocation	Region-ID	Dimension	This column was created in R by applying <code>as.interger(factor)</code> on Region column to create unique ID for available regions.
5	Growth-in-Percentage	FactTable	Growth-in-Percentage	Fact	Growth percentage values corresponding to each region and each year were obtained by using <code>substr()</code> function on the whole scrapped text from the Source Image

6 ETL Process

For this Data Warehousing project totally 11 data files have been used which are obtained from 5 different sources. Out of these 5, 4 are structured and 1 is unstructured. The details about the sources were given in section 2(Sources). All the structured data sources have been downloaded as .csv files and stored in a folder. This raw data is inserted/Updated into a staging area from where Dimensions and Fact table are created to facilitate a Data Warehouse model. The cleaning process were done using R and this process has been automated using SSIS by calling the R script inside SSIS ETL flow. The Processing of Cube and Dimensions have also been automated using SSIS. The Overall ETL Strategy using SSIS is shown in Figure:6



Figure 7: Energy - Star Schema

6.1 Extraction

Extraction is process of picking right data sources to include in to the Datawarehouse model. Six files from EnerData which gives information on Consumption, Production and Trade details of Natural gas and Oil providing information by region, country and Year wise were downloaded and stored as .csv file in local folder.

Two files from Carbon Atlas giving information on Co2 emisison on Natral gas and Oil were downloaded and stored in local folder as .csv format. Two files from Statista were downloaded which gives information on oil and gas Reserves which were then stored as .csv file in Local folder. One file from Kaggle which contains population information was downloaded and stored as .csv file in local folder. One file from WTO which is an Unstructured source of data is taken from online directly by R, which is then cleaned and processed in R automatically.

6.2 Transformation

The transformation process starts with converting the extracted data into the form as per the requirement. All data Transformation is done using R and this process is automated

using SSIS.

The files from EnerData is read from the local folder using R and is first cleaned to remove special characters, then the data was filtered to select only the required 44 countries and only for the period 2012 to 2016. The year data from row 2 is unlisted and made into column name. The columns were then transformed into rows using gather function so that consumption values are categorized by year wise for each country in rows. The same process was carried out for all 6 files from EnerData. Finally all 6 files were merged into one single file which forms one of the raw table (Raw-Energy). All the columns that is to be used in the fact table are casted into numeric value.

Similar process was done for the 2 datasets downloaded from the source Carbon atlas which were then merged into single Raw table (Raw-Co2Emission), and also for the dataset downloaded from the source Kaggle. All the columns that is to be used in the fact table are casted into numeric value.

The file from statista is read from the local folder and filtered to select only the required regions of interest and for the period 2012 to 2016. The second column which has Year information on rows is unlisted and made into column name. This was then transformed using gather function to categorize the reserve value by region and year wise in rows. Same process was done for one more file and two files were merged into one Raw Table (Raw-Reserve). All the columns that is to be used in the fact table are casted into numeric value.

The Unstructured data which is taken directly from online was scrapped using R. This was obtained as single text field. Each region and region wise GDP growth value was taken out individually using string operation. The special characters were removed and converted as numeric.

All the R code used for this project was attached in Appendix section.

6.3 Loading

Loading is the process of Insert/Update the cleaned and transformed data into the database, in this project SQL database server is used. The raw data loading process is shown in the below Figure:8

In this process, 4 raw tables were loaded with processed data as below.
The tool used for this Data warehousing project is SSIS toolbox.

- 1.Raw-Energy
- 2.Raw-Reserve
- 3.Raw-Co2-Emission
- 4.Raw-GDP-Growth
- 5.Raw-Population

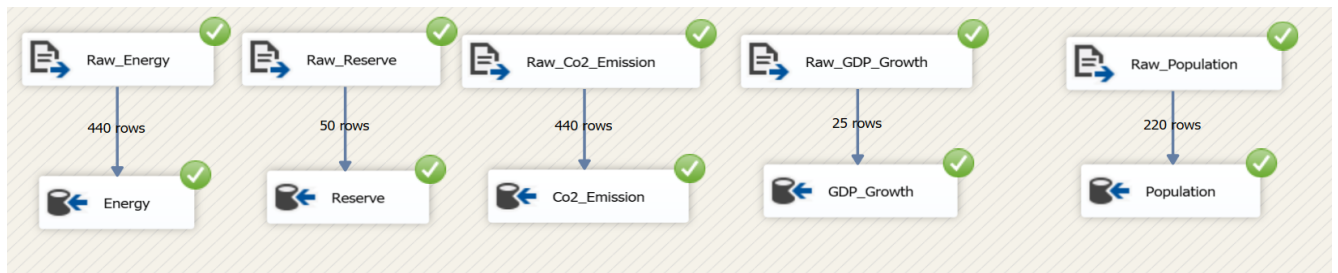


Figure 8: Raw Data Loading

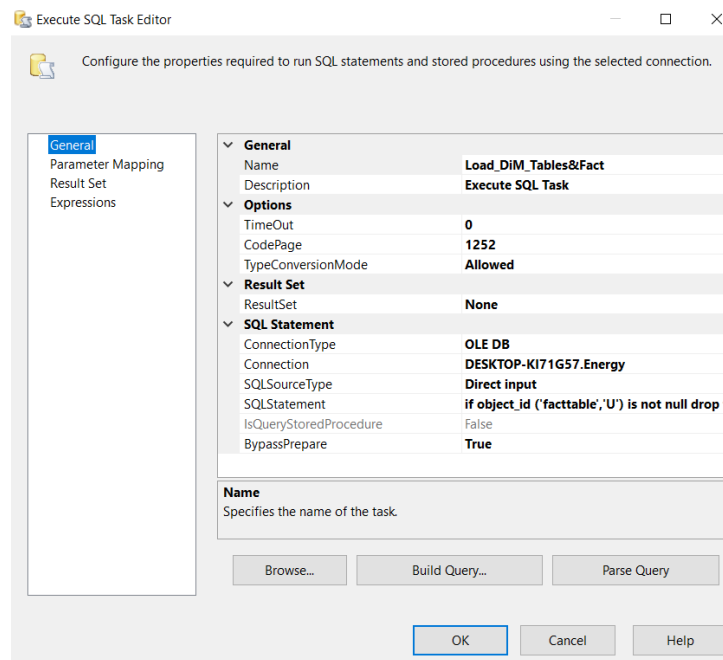


Figure 9: Dimension and Fact Table Loading

Once the Dimension tables and Fact tables were created, the tables are loaded with data according to the Warehouse model and business requirement. The SQL queries has been used to load the data into the tables. Related SQL queries can be found in the Appendix section at the end.

This process was done by using SSIS SQL task function. The below Dimension tables were loaded successfully.

- 1.Dim-Fuel-Type
- 2.Dim-Location
- 3.Dim-Time
- 4.FactTable

6.4 Cube Deployment

After the Dimension and Fact table are loaded with data, it must be processed and deployed. This part was done using SSAS package. Once the cube is deployed, the dimensions and fact values will be linked, and a relationship is established. Now the warehouse can be queried effectively using drill downs in cube browser. The Dimension and Cube processing is automated here by using SSIS.

7 Application

With all the technical aspects of the model were discussed above now the data warehouse model is ready to query and the business requirements can be analyzed. The three requirements(see section 1) has been staged and visualized using Tableau.

7.1 BI Query 1: Comparing the consumption of Natural gas and Oil globally and the emission of Co2 by each fuel source year wise and check for a correlation.

For this query, 2 different data sources have been used, Natural Gas and Oil consumption worldwide data which was taken from EnerData and Co2 emitted by Natural Gas and Oil which was taken from Carbon Atlas.

We could see that there is a correlation between the consumption of Gas and Oil sources to the emission of Co2 in each country and each year. This was illustrated in the below picture.

So from this diagram we can say that in order to reduce the emission of Co2 worldwide, alternative source of energy can be used instead of Gas and Oil for Domestic consumption and Industrial purposes.

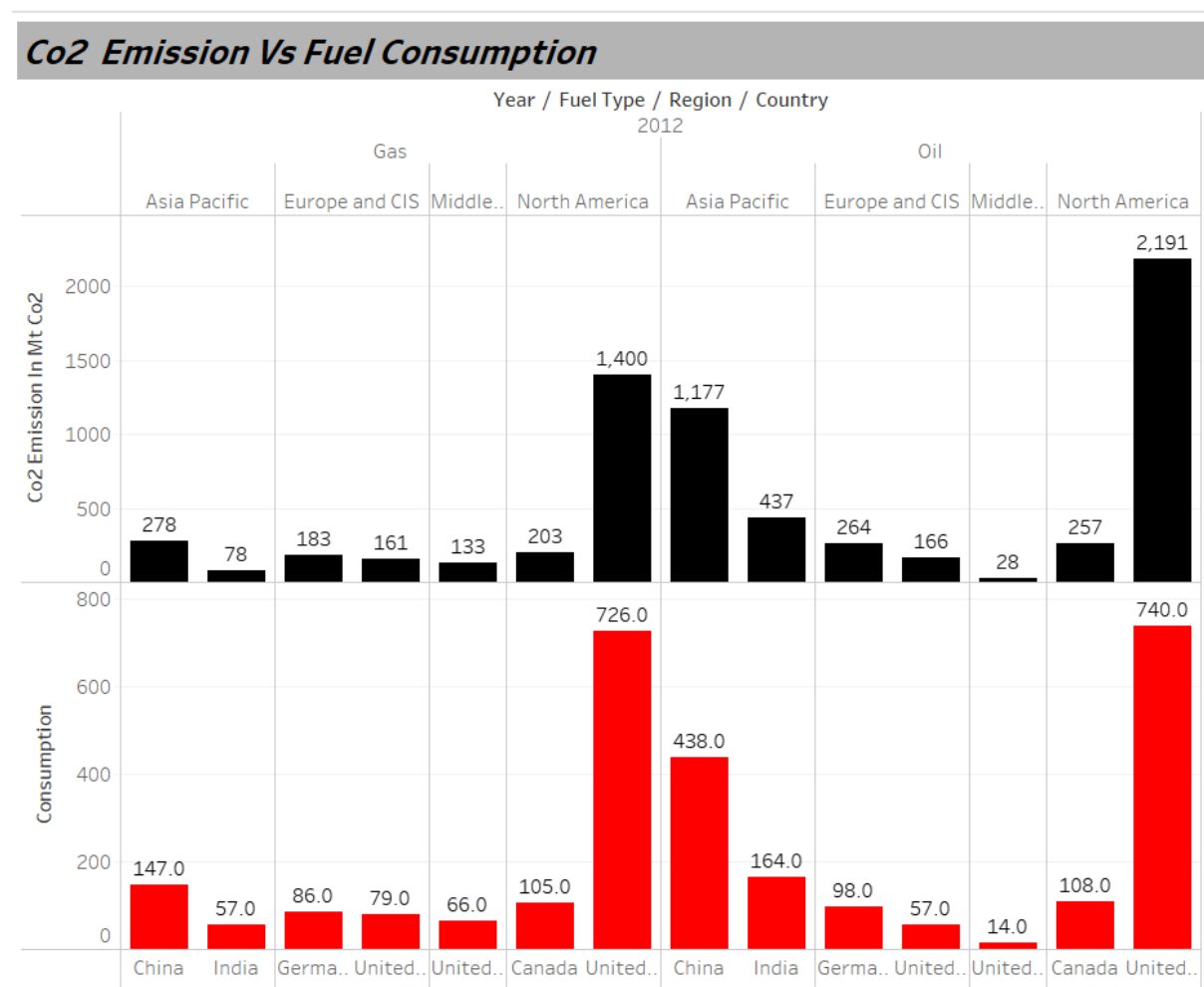


Figure 10: Results for BI Query 1

7.2 BI Query 2: Comparing the production of Natural gas and Oil worldwide based on resources available globally and check for a correlation.

For this query, 2 sources of data was used, one from EnerData giving information about Production of Natural Gas and Oil worldwide and another source from Statista providing information about percentage of Natural gas and Oil reserves available worldwide

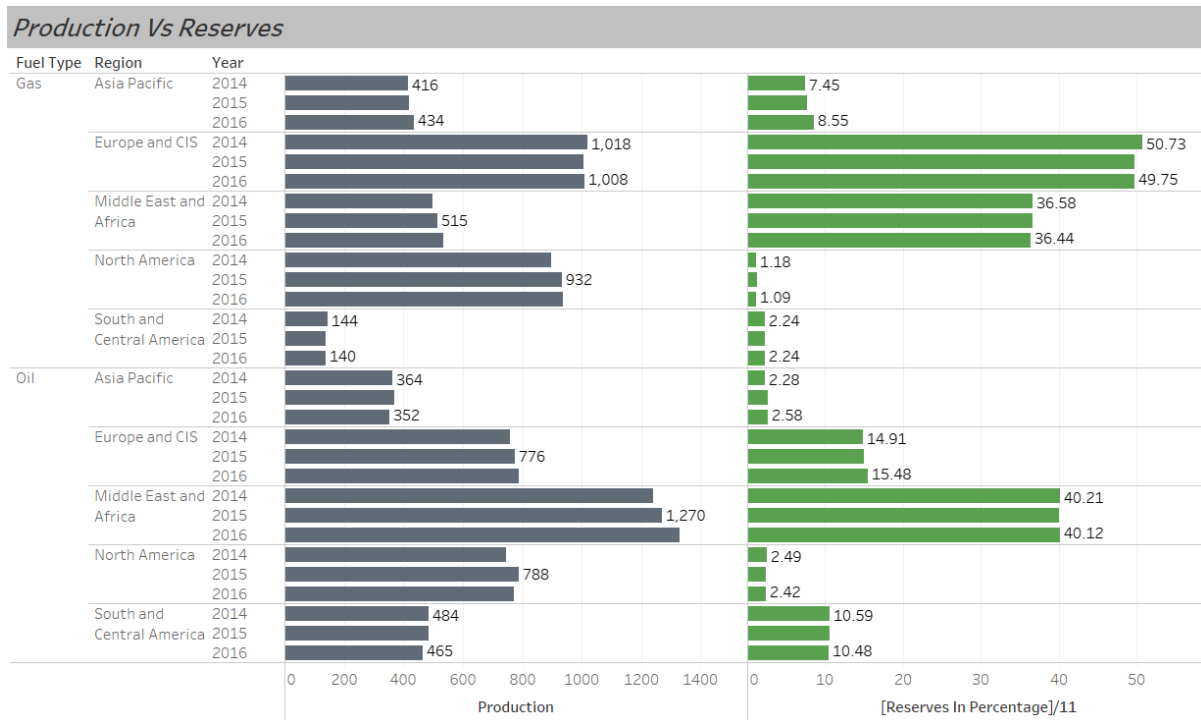


Figure 11: Results for BI Query 2

We could see that the production in a region is mainly based on the reserves available in some region like Europe and Middle East, apart from this there is not much significant co-relation between those two parameters. For example North America has greater reserves of Oil and Gas but the production is comparatively low.

7.3 BI Query 3: Comparing the Trade of Natural gas, Oil and Population worldwide and its impact on GDP Growth and check for a correlation among them.

For this Query, 2 sources of data was used, one from EnerData giving information about Trade of Natural Gas and Oil worldwide and another source which scrapped from a image from World Trade Organization website, this source gives information regarding GDP growth percentage each year.

Here, there is no co-relation between the Trade value and the region wise GDP growth each year, this means that the trade of Oil and Gas is not a major contributor in the economy of a region.

Also the Graph clearly depicts that the production is not based on Population, as there is no linear relationship between these two parameters.

Trade Vs GDP Growth and Population

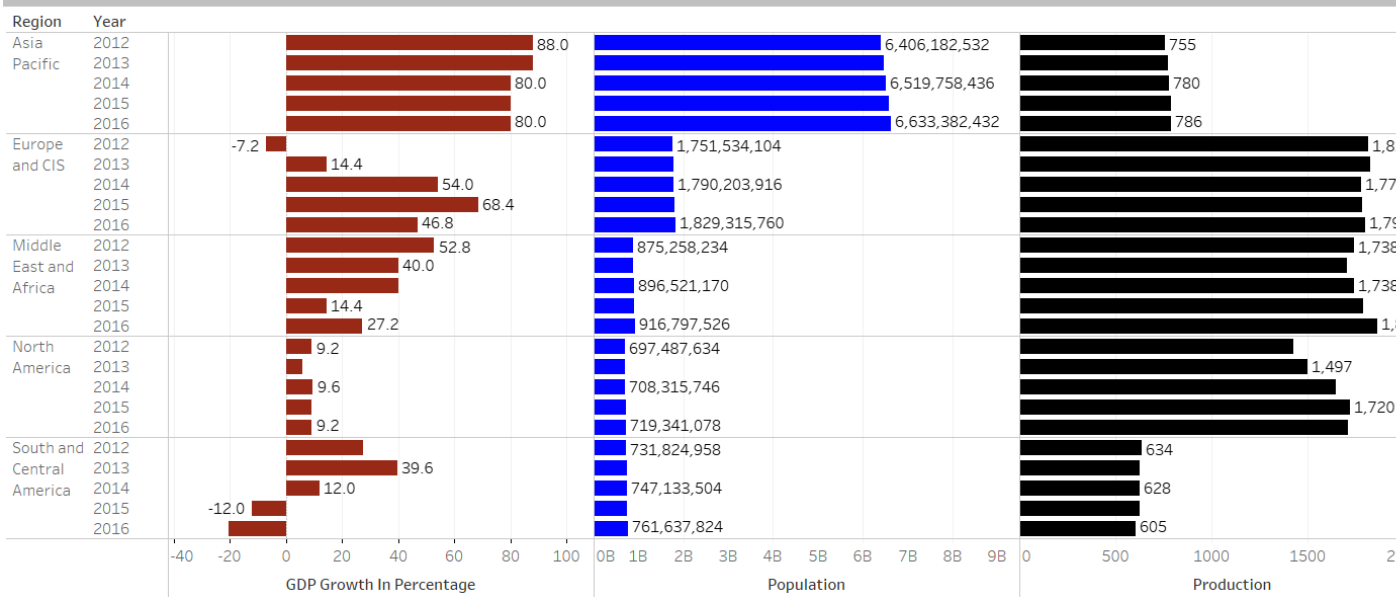


Figure 12: Results for BI Query 3

7.4 Discussion

In first query it is proved that the consumption of Natural Gas and Oil globally and Co2 emitted has correlation, that is the increase in usage of Gas and oil will increase the emission. Also, it is observed that emission is United States is more than other regions which is also discussed in (Nejat et al. 2015). This is in contradiction to what is being observed in Bildirici (2017) which concludes by saying that emission of Co2 is lowered by the usage of natural gas and renewable source of energy. This is supported by Dong et al. (2017) where it again emphasis on usage of Natural gas and its impact in reducing the Co2 emission except for United States and Europe.

In the second business requirement there is a correlation in reserves percentage and Production for Europe and apart from this there is not much significant correlation between Production and reserves available.

In the final query it is observed that there is no significant co-relation between GDP growth and Population and production of fossil fuels across globe. Bildirici (2017) a linear relation between Oil consumption and GDP growth is established in Middle East and Africa. Also, in Dong et al. (2017) a relationship is made between Oil and Gas emission and GDP Growth on BRICS countries. This paper also says that increase in consumption of Oil and Gas increase the emission by 0.165 to 0.26 percent.

8 Conclusion and Future Work

This project is made with a motive to analyze the relationship between Consumption, Production, Trade of Natural Gas with Oil Co2 Emission, Reserves available for Oil and Gas across the globe, Population and its impact on GDP growth worldwide. Many research projects were studied, and a data warehouse model is created. Data from various

relevant sources were gathered, cleaned and a model is built. The comparison between the various parameters are analyzed and visualized using Tableau and the results are discussed. It is observed that Co2 emission is mainly caused by the usage of Natural gas and Oil in various parts of the world. Except for Europe there is not much significance in the production of Oil and Gas based on the reserves available in specific regions. The production and consumption of Natural gas and oil do not have an impact on GDP growth globally.

Also, from this research it is observed that usage of Natural gas emits comparatively lesser amount of Co2 than Oil across the world.

This project does not support country wise drill down on reserves availability and GDP growth as the data could not be found and the future work of this project involves in attempting to obtain data on GDP growth for all countries based on the Consumption and Production of Oil and Natural Gas and to analyze the queries to more granular level. This project also aims to add more fossil fuels to the model in future and its effects and impacts on GDP growth and emission of Carbon dioxide.

References

Acheampong, A. O. (2018), 'Economic growth, co2 emissions and energy consumption: What causes what and where?', *Energy Economics* **74**, 677 – 692.

URL: <http://www.sciencedirect.com/science/article/pii/S014098831830272X>

Bildirici, M. (2017), 'Co2 emission, oil consumption and production, economic growth in menap countries: Ardl and anova methods', *International Journal of Oil, Gas and Coal Technology* **14**, 264.

Dong, K., Hochman, G., Zhang, Y., Sun, R., Li, H. & Liao, H. (2018), 'Co2 emissions, economic and population growth, and renewable energy: Empirical evidence across regions', *Energy Economics* **75**, 180 – 192.

URL: <http://www.sciencedirect.com/science/article/pii/S0140988318303256>

Dong, K., Sun, R. & Hochman, G. (2017), 'Do natural gas and renewable energy consumption lead to less co2 emission? empirical evidence from a panel of brics countries', *Energy* **141**, 1466 – 1478.

URL: <http://www.sciencedirect.com/science/article/pii/S0360544217319503>

Kimball, R., Ross, M. & Kimball, R. (2013), *The data warehouse toolkit : the definitive guide to dimensional modeling : [electronic book].*, Indianapolis, Ind. : John Wiley, 2013.

URL: <http://search.ebscohost.com/login.aspx?direct=trueAuthType=ip,cookie,shibdb=cat05743aAN=livescope=sitcustid=ncirlib>

Nejat, P., Jomehzadeh, F., Taheri, M. M., Gohari, M. & Majid, M. Z. A. (2015), 'A global review of energy consumption, co2 emissions and policy in the residential sector (with an overview of the top ten co2 emitting countries)', *Renewable and Sustainable Energy Reviews* **43**, 843 – 862.

URL: <http://www.sciencedirect.com/science/article/pii/S1364032114010053>

Appendix

Data Cleaning and Transformation for Raw-Energy

```
#Script-1 (Data Cleaning and Transformation for Raw-Energy)
  getwd()
  setwd("C:/Users/Pavan_Kumar/Desktop/DWBI/WorkingDirectory/")
  library(dplyr)
  library(tidyr)
  library(xml2)
  library(rvest)
  library(sqldf)

#Oil Production table formation.

Oil_Production <- read.csv("C:/Users/
_Pavan_Kumar/Desktop/DWBI/Energy/Crude_Oil_Production.csv")
Oil_Production <- Oil_Production[c(-1,-3:-7),]
colnames(Oil_Production) <- as.character(unlist(Oil_Production[1,]))
colnames(Oil_Production)[colnames(Oil_Production) == "1" ] <- 'Country'
Oil_Production <- Oil_Production[-1,]
#Grouping countries according to thier regions.
Oil_Production$Region <- ifelse(Oil_Production$Country
== 'Canada',"North_America",
ifelse(Oil_Production$Country == 'United_States',"North_America",
ifelse(Oil_Production$Country == 'Argentina',"South_and_Central_America",
ifelse(Oil_Production$Country == 'Brazil',"South_and_Central_America",
ifelse(Oil_Production$Country == 'Chile',"South_and_Central_America",
ifelse(Oil_Production$Country == 'Colombia',"South_and_Central_America",
ifelse(Oil_Production$Country == 'Mexico',"South_and_Central_America",
ifelse(Oil_Production$Country == 'Venezuela',"South_and_Central_America",
ifelse(Oil_Production$Country == 'China',"Asia_Pacific",
ifelse(Oil_Production$Country == 'India',"Asia_Pacific",
ifelse(Oil_Production$Country == 'Indonesia',"Asia_Pacific",
ifelse(Oil_Production$Country == 'Japan',"Asia_Pacific",
ifelse(Oil_Production$Country == 'Malaysia',"Asia_Pacific",
ifelse(Oil_Production$Country == 'South_Korea',"Asia_Pacific",
ifelse(Oil_Production$Country == 'Taiwan',"Asia_Pacific",
ifelse(Oil_Production$Country == 'Thailand',"Asia_Pacific",
ifelse(Oil_Production$Country == 'Australia',"Asia_Pacific",
ifelse(Oil_Production$Country == 'New_Zealand',"Asia_Pacific",
ifelse(Oil_Production$Country == 'South_Africa',"Middle_East_and_Africa",
ifelse(Oil_Production$Country == 'Algeria',"Middle_East_and_Africa",
ifelse(Oil_Production$Country == 'Egypt',"Middle_East_and_Africa",
ifelse(Oil_Production$Country == 'Nigeria',"Middle_East_and_Africa",
ifelse(Oil_Production$Country == 'Iran',"Middle_East_and_Africa",
ifelse(Oil_Production$Country == 'Kuwait',"Middle_East_and_Africa",
ifelse(Oil_Production$Country == 'Saudi_Arabia',"Middle_East_and_Africa",
ifelse(Oil_Production$Country == 'United_Arab_Emirates',"Middle_East
and_Africa","Europe_and_CIS")))))))
```



```

rownames(Oil_Production) <- c(1:nrow(Oil_Production))
Oil_Production
<- Oil_Production[c(-1,-16,-21,-22,-25,-32,-41,-44,-49,-54,-55),]
Oil_Production <- Oil_Production[,c(1,24:28,32)]
Oil_Production <- Oil_Production[c(7,1:6)]

#Transforming table structure to facilitate drill down

Oil_Production <- Oil_Production
%>% gather(Year,Production,'2012','2013','2014','2015','2016')
Oil_Production$Fuel_Type <- rep("Oil",nrow(Oil_Production))
Oil_Production <- Oil_Production[c(3,5,1,2,4)]
write.csv(Oil_Production,file = "Oil_Production.csv",row.names = FALSE)

#Oil Trade table formation.

Oil_Trade <- read.csv("C:/Users/
Pavan_Kumar/Desktop/DWBI/Energy/Crude_Oil_Trade.csv")
Oil_Trade <- Oil_Trade[c(-1,-3:-6),]
colnames(Oil_Trade) <- as.character(unlist(Oil_Trade[1,]))
colnames(Oil_Trade)[colnames(Oil_Trade) == "1" ] <- 'Country'
Oil_Trade <- Oil_Trade[-1,]
Oil_Trade$Region <- ifelse(Oil_Trade$Country ==
'Canada',"North_America",
ifelse(Oil_Trade$Country == 'United_States',"North_America",
ifelse(Oil_Trade$Country == 'Argentina',"South_and_Central_America",
ifelse(Oil_Trade$Country == 'Brazil',"South_and_Central_America",
ifelse(Oil_Trade$Country == 'Chile',"South_and_Central_America",
ifelse(Oil_Trade$Country == 'Colombia',"South_and_Central_America",
ifelse(Oil_Trade$Country == 'Mexico',"South_and_Central_America",
ifelse(Oil_Trade$Country == 'Venezuela',"South_and_Central_America",
ifelse(Oil_Trade$Country == 'China',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'India',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'Indonesia',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'Japan',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'Malaysia',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'South_Korea',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'Taiwan',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'Thailand',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'Australia',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'New_Zealand',"Asia_Pacific",
ifelse(Oil_Trade$Country == 'South_Africa',"Middle_East_and_Africa",
ifelse(Oil_Trade$Country == 'Algeria',"Middle_East_and_Africa",
ifelse(Oil_Trade$Country == 'Egypt',"Middle_East_and_Africa",
ifelse(Oil_Trade$Country == 'Nigeria',"Middle_East_and_Africa",
ifelse(Oil_Trade$Country == 'Iran',"Middle_East_and_Africa",
ifelse(Oil_Trade$Country == 'Kuwait',"Middle_East_and_Africa",
ifelse(Oil_Trade$Country == 'Saudi_Arabia',"Middle_East_and_Africa",
ifelse(Oil_Trade$Country == 'United_Arab_Emirates',
"Middle_East_and_Africa","Europe_and_CIS")))))))

```

```

rownames(Oil_Trade) <- c(1:nrow(Oil_Trade))
Oil_Trade <- Oil_Trade[c(-1,-16,-21,-22,-25,-32,-41,-44,-49,-54,-55),]
Oil_Trade <- Oil_Trade[,c(1,24:28,32)]
Oil_Trade <- Oil_Trade[c(7,1:6)]
Oil_Trade <- Oil_Trade
%>% gather(Year,Trade,'2012','2013','2014','2015','2016')
Oil_Trade$Fuel_Type <- rep("Oil",nrow(Oil_Trade))
Oil_Trade <- Oil_Trade[c(3,5,1,2,4)]
write.csv(Oil_Trade,file = "Oil_Trade.csv",row.names = FALSE)

Oil_Consumption <- read.csv("C:/Users/
Pavan_Kumar/Desktop/DWBI/Energy/Crude_Oil_Consumption.csv")
Oil_Consumption <- Oil_Consumption[c(-1,-3:-7),]
colnames(Oil_Consumption) <- as.character(unlist(Oil_Consumption[1,]))
colnames(Oil_Consumption)[colnames(Oil_Consumption) == "1" ] <- 'Country'
Oil_Consumption <- Oil_Consumption[-1,]
Oil_Consumption$Region <- ifelse(Oil_Consumption$Country
== 'Canada',"North_America",
ifelse(Oil_Consumption$Country == 'United_States',"North_America",
ifelse(Oil_Consumption$Country == 'Argentina',"South_and_Central_America",
ifelse(Oil_Consumption$Country == 'Brazil',"South_and_Central_America",
ifelse(Oil_Consumption$Country == 'Chile',"South_and_Central_America",
ifelse(Oil_Consumption$Country == 'Colombia',"South_and_Central_America",
ifelse(Oil_Consumption$Country == 'Mexico',"South_and_Central_America",
ifelse(Oil_Consumption$Country == 'Venezuela',"South_and_Central_America",
ifelse(Oil_Consumption$Country == 'China',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'India',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'Indonesia',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'Japan',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'Malaysia',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'South_Korea',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'Taiwan',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'Thailand',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'Australia',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'New_Zealand',"Asia_Pacific",
ifelse(Oil_Consumption$Country == 'South_Africa',"Middle_East_and_Africa",
ifelse(Oil_Consumption$Country == 'Algeria',"Middle_East_and_Africa",
ifelse(Oil_Consumption$Country == 'Egypt',"Middle_East_and_Africa",
ifelse(Oil_Consumption$Country == 'Nigeria',"Middle_East_and_Africa",
ifelse(Oil_Consumption$Country == 'Iran',"Middle_East_and_Africa",

ifelse(Oil_Consumption$Country
== 'Saudi_Arabia',"Middle_East_and_Africa",
ifelse(Oil_Consumption$Country
== 'United_Arab_Emirates',"Middle_East_and_Africa","Europe_and_CIS"))))))))
rownames(Oil_Consumption) <- c(1:nrow(Oil_Consumption))
Oil_Consumption
<- Oil_Consumption[c(-1,-16,-21,-22,-25,-32,-41,-44,-49,-54,-55),]
Oil_Consumption <- Oil_Consumption[,c(1,24:28,32)]

```

```

Oil_Consumption <- Oil_Consumption[c(7,1:6)]
Oil_Consumption <- Oil_Consumption
%>% gather(Year,Consumption,'2012','2013','2014','2015','2016')
Oil_Consumption$Fuel_Type <- rep("Oil",nrow(Oil_Consumption))
Oil_Consumption <- Oil_Consumption[c(3,5,1,2,4)]
write.csv(Oil_Consumption,file = "Oil_Consumption.csv",row.names = FALSE)

#-- Gas-----

Gas_Production <- read.csv("C:/Users/
Pavan_Kumar/Desktop/DWBI/Energy/Natural_Gas_Production.csv")
Gas_Production <- Gas_Production[c(-1,-3:-7),]
colnames(Gas_Production) <- as.character(unlist(Gas_Production[1,]))
colnames(Gas_Production)[colnames(Gas_Production) == "1" ] <- 'Country'
Gas_Production <- Gas_Production[-1,]
Gas_Production$Region <- ifelse(Gas_Production$Country ==
'Canada','North_America',
ifelse(Gas_Production$Country == 'United_States','North_America',
ifelse(Gas_Production$Country == 'Argentina','South_and_Central_America',
ifelse(Gas_Production$Country == 'Brazil','South_and_Central_America',
ifelse(Gas_Production$Country == 'Chile','South_and_Central_America',
ifelse(Gas_Production$Country == 'Colombia','South_and_Central_America',
ifelse(Gas_Production$Country == 'Mexico','South_and_Central_America',
ifelse(Gas_Production$Country == 'Venezuela','South_and_Central_America',
ifelse(Gas_Production$Country == 'China','Asia_Pacific',
ifelse(Gas_Production$Country == 'India','Asia_Pacific',
ifelse(Gas_Production$Country == 'Indonesia','Asia_Pacific',
ifelse(Gas_Production$Country == 'Japan','Asia_Pacific',
ifelse(Gas_Production$Country == 'Malaysia','Asia_Pacific',
ifelse(Gas_Production$Country == 'South_Korea','Asia_Pacific',
ifelse(Gas_Production$Country == 'Taiwan','Asia_Pacific',
ifelse(Gas_Production$Country == 'Thailand','Asia_Pacific',
ifelse(Gas_Production$Country == 'Australia','Asia_Pacific',
ifelse(Gas_Production$Country == 'New_Zealand','Asia_Pacific',
ifelse(Gas_Production$Country == 'South_Africa','Middle_East_and_Africa',
ifelse(Gas_Production$Country == 'Algeria','Middle_East_and_Africa',
ifelse(Gas_Production$Country == 'Egypt','Middle_East_and_Africa',
ifelse(Gas_Production$Country == 'Nigeria','Middle_East_and_Africa',
ifelse(Gas_Production$Country == 'Iran','Middle_East_and_Africa',
ifelse(Gas_Production$Country == 'Kuwait','Middle_East_and_Africa',
ifelse(Gas_Production$Country == 'Saudi_Arabia','Middle_East
and_Africa',
ifelse(Gas_Production$Country == 'United_Arab_Emirates','Middle_East
and_Africa',"Europe_and_CIS"))))))))))))))))))))
rownames(Gas_Production) <- c(1:nrow(Gas_Production))
Gas_Production
<- Gas_Production[c(-1,-16,-21,-22,-25,-32,-41,-44,-49,-54,-55),]
Gas_Production <- Gas_Production[,c(1,24:28,32)]
Gas_Production <- Gas_Production[c(7,1:6)]

```

```

Gas_Production <- Gas_Production
%>% gather(Year,Production,'2012','2013','2014','2015','2016')
Gas_Production$Fuel_Type <- rep("Gas",nrow(Gas_Production))
Gas_Production <- Gas_Production[c(3,5,1,2,4)]
write.csv(Gas_Production,file = "Gas_Production.csv",row.names = FALSE)

Gas_Trade <- read.csv("C:/Users/
Pavan_Kumar/Desktop/DWBI/Energy/Natural_Gas_Trade.csv")
Gas_Trade <- Gas_Trade[c(-1,-3:-6),]
colnames(Gas_Trade) <- as.character(unlist(Gas_Trade[1,]))
colnames(Gas_Trade)[colnames(Gas_Trade) == "1" ] <- 'Country'
Gas_Trade <- Gas_Trade[-1,]
Gas_Trade$Region <- ifelse(Gas_Trade$Country == 'Canada',"North_America",
ifelse(Gas_Trade$Country == 'United_States',"North_America",
ifelse(Gas_Trade$Country == 'Argentina',"South_and_Central_America",
ifelse(Gas_Trade$Country == 'Brazil',"South_and_Central_America",
ifelse(Gas_Trade$Country == 'Chile',"South_and_Central_America",
ifelse(Gas_Trade$Country == 'Colombia',"South_and_Central_America",
ifelse(Gas_Trade$Country == 'Mexico',"South_and_Central_America",
ifelse(Gas_Trade$Country == 'Venezuela',"South_and_Central_America",
ifelse(Gas_Trade$Country == 'China',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'India',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'Indonesia',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'Japan',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'Malaysia',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'South_Korea',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'Taiwan',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'Thailand',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'Australia',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'New_Zealand',"Asia_Pacific",
ifelse(Gas_Trade$Country == 'South_Africa',"Middle_East_and_Africa",
ifelse(Gas_Trade$Country == 'Algeria',"Middle_East_and_Africa",
ifelse(Gas_Trade$Country == 'Egypt',"Middle_East_and_Africa",
ifelse(Gas_Trade$Country == 'Nigeria',"Middle_East_and_Africa",
ifelse(Gas_Trade$Country == 'Iran',"Middle_East_and_Africa",
ifelse(Gas_Trade$Country == 'Kuwait',"Middle_East_and_Africa",
ifelse(Gas_Trade$Country == 'Saudi_Arabia',"Middle_East_and_Africa",
ifelse(Gas_Trade$Country == 'United_Arab_Emirates',"Middle_East
and_Africa","Europe_and_CIS")))))))
rownames(Gas_Trade) <- c(1:nrow(Gas_Trade))
Gas_Trade <- Gas_Trade[c(-1,-16,-21,-22,-25,-32,-41,-44,-49,-54,-55),]
Gas_Trade <- Gas_Trade[,c(1,24:28,32)]
Gas_Trade <- Gas_Trade[c(7,1:6)]
Gas_Trade <- Gas_Trade %>% gather
(Year,Trade,'2012','2013','2014','2015','2016')
Gas_Trade$Fuel_Type <- rep("Gas",nrow(Gas_Trade))
Gas_Trade <- Gas_Trade[c(3,5,1,2,4)]
write.csv(Gas_Trade,file = "Gas_Trade.csv",row.names = FALSE)

```

```

Gas_Consumption <- read.csv("C:/Users/
_Pavan_Kumar/Desktop/DWBI/Energy/Natural_Gas_Consumption.csv")
Gas_Consumption <- Gas_Consumption[c(-1,-3:-7),]
colnames(Gas_Consumption) <- as.character(unlist(Gas_Consumption[1,]))
colnames(Gas_Consumption)[colnames(Gas_Consumption) == "1" ] <- 'Country'
Gas_Consumption <- Gas_Consumption[-1,]
Gas_Consumption$Region <- ifelse(Gas_Consumption$Country
== 'Canada',"North_America",
ifelse(Gas_Consumption$Country == 'United_States',"North_America",
ifelse(Gas_Consumption$Country == 'Argentina',"South_and_Central_America",
ifelse(Gas_Consumption$Country == 'Brazil',"South_and_Central_America",
ifelse(Gas_Consumption$Country == 'Chile',"South_and_Central_America",
ifelse(Gas_Consumption$Country == 'Colombia',"South_and_Central_America",
ifelse(Gas_Consumption$Country == 'Mexico',"South_and_Central_America",
ifelse(Gas_Consumption$Country == 'Venezuela',"South_and_Central_America",
ifelse(Gas_Consumption$Country == 'China',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'India',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'Indonesia',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'Japan',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'Malaysia',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'South_Korea',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'Taiwan',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'Thailand',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'Australia',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'New_Zealand',"Asia_Pacific",
ifelse(Gas_Consumption$Country == 'South_Africa',"Middle_East_and_Africa",
ifelse(Gas_Consumption$Country == 'Algeria',"Middle_East_and_Africa",
ifelse(Gas_Consumption$Country == 'Egypt',"Middle_East_and_Africa",
ifelse(Gas_Consumption$Country == 'Nigeria',"Middle_East_and_Africa",
ifelse(Gas_Consumption$Country == 'Iran',"Middle_East_and_Africa",
ifelse(Gas_Consumption$Country == 'Kuwait',"Middle_East_and_Africa",
ifelse(Gas_Consumption$Country == 'Saudi_Arabia',"Middle_East_and_Africa",

rownames(Gas_Consumption) <- c(1:nrow(Gas_Consumption))
Gas_Consumption <- Gas_Consumption
[c(-1,-16,-21,-22,-25,-32,-41,-44,-49,-54,-55),]
Gas_Consumption <- Gas_Consumption[,c(1,24:28,32)]
Gas_Consumption <- Gas_Consumption[c(7,1:6)]
Gas_Consumption <- Gas_Consumption %>% gather(Year,Consumption,
'2012','2013','2014','2015','2016')
Gas_Consumption$Fuel_Type <- rep("Gas",nrow(Gas_Consumption))
Gas_Consumption <- Gas_Consumption[c(3,5,1,2,4)]
write.csv(Gas_Consumption,file = "Gas_Consumption.csv",row.names = FALSE)

#- Merging Oil and Gas Category wise to form a single table

Energy_Consumption <- rbind(Oil_Consumption,Gas_Consumption)
Energy_Production <- rbind(Oil_Production,Gas_Production)
Energy_Trade <- rbind(Oil_Trade,Gas_Trade)

```



```

Energy <- cbind.data.frame(Energy_Consumption ,
Energy_Production$Production,Energy_Trade$Trade)
colnames(Energy) <- c("Year","Fuel_Type","Region","Country","Consumption",
"Production","Trade")

Energy$Fuel_ID <- as.integer(factor(Energy$Fuel_Type))
Energy$Region_ID <- as.integer(factor(Energy$Region))
Energy$Country_ID <- as.integer(factor(Energy$Country))
Energy <- Energy[order(Energy$Region),]
Energy <- Energy[c(1,8,2,9,3,10,4,5,6,7)]
write.csv(Energy,file = "Energy.csv",row.names = FALSE)

```

Data Cleaning and Transformation for Raw Co2 Emisison

```

setwd("C:/Users/Pavan_Kumar/Desktop/DWBI/WorkingDirectory/")
library(dplyr)
library(tidyr)
library(xml2)
library(rvest)
library(sqldf)

Co2_Oil <- read.csv("C:/Users/Pavan_Kumar/Desktop/DWBI/Energy/Oil_CO2.csv")
Co2_Oil[1,1] <- "Year"
Co2_Oil <- Co2_Oil[,c(1,4,9,12,20,29,37,42,43,44,54,61,71,77,93,94,95,
99,101,103,106,121,129,141,143,146,149,159,160,164,165,174,184,185
,186,191,194,197,203,208,209,210,211,213,215)]
colnames(Co2_Oil) colnames(Co2_Oil) <-
c("Year","Algeria","Argentina",
"Australia","Belgium","Brazil","Canada","Chile",
"China","Colombia","Czech_Republic","Egypt",
"France","Germany","India","Indonesia","Iran",
"Italy","Japan","Kazakhstan","Kuwait",
"Malaysia","Mexico","Netherlands","New_Zealand","Nigeria","Norway","Poland",
"Portugal","Romania","Russian_Federation",
,"Saudi_Arabia","South_Africa",
"South_Korea","Spain","Sweden","Taiwan","Thailand","Turkey","Ukraine",
"United_Arab_Emirates", "United_Kingdom",
"United_States","Uzbekistan", "Venezuela")
Co2_Oil <- Co2_Oil[-1:-3,]
Co2_Oil <- Co2_Oil %>%
(Country,Co2_Emission_in_mtCo2, Algeria,Argentina,
Australia,Belgium,Brazil,Canada,Chile,China
,Colombia,Czech Republic,Egypt,
France,Germany,India,Indonesia,Iran,Italy
,Japan,Kazakhstan,Kuwait,
Malaysia,Mexico,Netherlands,New Zealand
,Nigeria,Norway,Poland,Portugal,
Romania,Russian Federation,Saudi Arabia,South Africa,
South Korea,Spain,Sweden,Taiwan,Thailand,Turkey,Ukraine,
United Arab Emirates, United Kingdom,

```

```

United States,Uzbekistan, Venezuela , -Year)
rownames(Co2_Oil) <- c(1:nrow(Co2_Oil))
Co2_Oil$Region ifelse(Co2_Oil$Country == 'Canada',"North_America",
  ifelse(Co2_Oil$Country == 'United_States',"North_America",
    ifelse(Co2_Oil$Country == 'Argentina',"South_and_Central_America",
      ifelse(Co2_Oil$Country == 'Brazil',"South_and_Central_America",
        ifelse(Co2_Oil$Country == 'Chile',"South_and_Central_America",
          ifelse(Co2_Oil$Country == 'Colombia',"South_and_Central_America",
            ifelse(Co2_Oil$Country == 'Mexico',"South_and_Central_America",
              ifelse(Co2_Oil$Country == 'Venezuela',"South_and_Central_America",
                ifelse(Co2_Oil$Country == 'China',"Asia_Pacific",
                  ifelse(Co2_Oil$Country == 'India',"Asia_Pacific",
                    ifelse(Co2_Oil$Country == 'Indonesia',"Asia_Pacific",
                      ifelse(Co2_Oil$Country == 'Japan',"Asia_Pacific",
                        ifelse(Co2_Oil$Country == 'Malaysia',"Asia_Pacific",
                          ifelse(Co2_Oil$Country == 'South_Korea',"Asia_Pacific",
                            ifelse(Co2_Oil$Country == 'Taiwan',"Asia_Pacific",
                              ifelse(Co2_Oil$Country == 'Thailand',"Asia_Pacific",
                                ifelse(Co2_Oil$Country == 'Australia',"Asia_Pacific",
                                  ifelse(Co2_Oil$Country == 'New_Zealand',"Asia_Pacific",
                                    ifelse(Co2_Oil$Country == 'South_Africa',"Middle_East_and_Africa",
                                      ifelse(Co2_Oil$Country == 'Algeria',"Middle_East_and_Africa",
                                        ifelse(Co2_Oil$Country == 'Egypt',"Middle_East_and_Africa",
                                          ifelse(Co2_Oil$Country == 'Nigeria',"Middle_East_and_Africa",
                                            ifelse(Co2_Oil$Country == 'Iran',"Middle_East_and_Africa",
                                              ifelse(Co2_Oil$Country == 'Kuwait',"Middle_East_and_Africa",
                                                ifelse(Co2_Oil$Country == 'Saudi_Arabia',"Middle_East_and_Africa",
                                                  ifelse(Co2_Oil$Country == 'United_Arab_Emirates',"Middle_East_and_Africa",
                                                    "Europe_and_CIS")))))))))))))))))))))))))))
Co2_Oil <- Co2_Oil[c(1,4,2,3)]
write.csv(Co2_Oil,file = "Co2_Oil.csv",row.names = FALSE)

```

#-----Gas-----

```

Co2_Gas <- read.csv("C:/Users/Pavan_Kumar/Desktop/DWBI/Energy/Gas_CO2.csv")
Co2_Gas[1,1] <- "Year"
Co2_Gas <- Co2_Gas[,c(1,4,9,12,20,29,37,42,43,44,54,61,71,77,93,94,95,99,
101,103,106,121,129,141,143,146,149,159,160,164,165,174,184,185,186,191,
194,197,203,208,209,210,211,213,215)]
colnames(Co2_Gas) <-
c("Year","Algeria","Argentina",
"Australia","Belgium","Brazil","Canada","Chile",
"China","Colombia",
"Czech_Republic","Egypt",
"France","Germany","India","Indonesia","Iran","Italy","Japan",
"Kazakhstan","Kuwait",

```

```

"Malaysia","Mexico","Netherlands","New_Zealand","Nigeria","Norway","Poland",
"Romania","Russian_Federation","Saudi_Arabia","South_Africa",
"South_Korea","Spain","Sweden","Taiwan","Thailand","Turkey","Ukraine",
"United_Arab_Emirates","United_Kingdom",
"United_States","Uzbekistan","Venezuela")Co2_Gas <- Co2_Gas[-1:-3,]
Co2_Gas <- Co2_Gas %>% gather
(Country,Co2_Emission_in_mtCo2, Algeria,Argentina,Australia,
Belgium,Brazil,
Canada,Chile,China,Colombia,Czech Republic,
Egypt,France,Germany,India,Indonesia,Iran,Italy,Japan,Kazakhstan,
Kuwait,Malaysia,Mexico,Netherlands,New Zealand,Nigeria,
Norway,Poland,Portugal,Romania,Russian Federation,
Saudi Arabia,South Africa,
South Korea,Spain,Sweden,Taiwan,Thailand,Turkey,Ukraine,
United Arab Emirates, United Kingdom,
United States,Uzbekistan, Venezuela ,-Year)
rownames(Co2_Gas) <- c(1:nrow(Co2_Gas))
Co2_Gas$Region <- ifelse(Co2_Gas$Country == 'Canada',"North_America",
ifelse(Co2_Gas$Country == 'United_States',"North_America",
ifelse(Co2_Gas$Country == 'Argentina',"South_and_Central_America",
ifelse(Co2_Gas$Country == 'Brazil',"South_and_Central_America",
ifelse(Co2_Gas$Country == 'Chile',"South_and_Central_America",
ifelse(Co2_Gas$Country == 'Colombia',"South_and_Central_America",
ifelse(Co2_Gas$Country == 'Mexico',"South_and_Central_America",
ifelse(Co2_Gas$Country == 'Venezuela',"South_and_Central_America",
ifelse(Co2_Gas$Country == 'China',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'India',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'Indonesia',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'Japan',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'Malaysia',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'South_Korea',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'Taiwan',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'Thailand',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'Australia',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'New_Zealand',"Asia_Pacific",
ifelse(Co2_Gas$Country == 'South_Africa',"Middle_East_and_Africa",
ifelse(Co2_Gas$Country == 'Algeria',"Middle_East_and_Africa",
ifelse(Co2_Gas$Country == 'Egypt',"Middle_East_and_Africa",
ifelse(Co2_Gas$Country == 'Nigeria',"Middle_East_and_Africa",
ifelse(Co2_Gas$Country == 'Iran',"Middle_East_and_Africa",
ifelse(Co2_Gas$Country == 'Kuwait',"Middle_East_and_Africa",
ifelse(Co2_Gas$Country == 'Saudi_Arabia',"Middle_East_and_Africa",
ifelse(Co2_Gas$Country == 'United_Arab_Emirates',"Middle_East_and_Africa",
"Europe_and_CIS"))))))))))))))))))))))))))))
Co2_Gas <- Co2_Gas[c(1,4,2,3)]
write.csv(Co2_Gas,file = "Co2_Gas.csv",row.names = FALSE)
Co2_Gas <- Co2_Gas[order(Co2_Gas$Region),]
Co2_Oil <- Co2_Oil[order(Co2_Oil$Region),]
Co2_Oil$Co2_Emission_in_mtCo2
<- format(round(as.numeric(Co2_Oil$Co2_Emission_in_mtCo2), 2), nsmall = 2)
Co2_Gas$Co2_Emission_in_mtCo2

```



```

<- format(round(as.numeric(Co2_Gas$Co2_Emission_in_mtCo2), 2), nsmall = 2)
Co2_Gas$Fuel_Type <- rep("Gas",nrow(Co2_Gas))
Co2_Oil$Fuel_Type <- rep("Oil",nrow(Co2_Oil))
Co2_Emission <- rbind(Co2_Gas,Co2_Oil)
Co2_Emission$Region_ID <- as.integer(factor(Co2_Emission$Region))
Co2_Emission$Country_ID <- as.integer(factor(Co2_Emission$Country))
Co2_Emission$Fuel_ID <- as.integer(factor(Co2_Emission$Fuel_Type))
Co2_Emission <- Co2_Emission[c(1,8,5,6,2,7,3,4)]
write.csv(Co2_Emission,file = "Co2_Emission.csv",row.names = FALSE)

```

Data Cleaning and Transformation for Raw Reserve

```

setwd("C:/Users/Pavan_Kumar/Desktop/DWBI/WorkingDirectory/")
library(dplyr)
library(tidyr)
library(xml2)
library(rvest)
library(sqldf)

```

```

Gas_Reserve <-
read.csv("C:/Users/PavanKumar/Desktop/DWBI/Energy/
Natural_Gas_Reserves.csv")
colnames(Gas_Reserve) <- as.character(unlist(Gas_Reserve[2,]))
Gas_Reserve <- Gas_Reserve[c(-1:-2),c(-2,-3,-9,-10)]
colnames(Gas_Reserve)[colnames(Gas_Reserve) == "1" ]
<- 'Region'
Gas_Reserve <- Gas_Reserve %>% gather(Year,Reserves_in_Percentage,
'2012','2013','2014','2015','2016',-Region)
write.csv(Gas_Reserve,file = "Gas_Reserve.csv", row.names = FALSE)

```

```

Oil_Reserve <- read.csv("C:/Users/
Pavan_Kumar/Desktop/DWBI/Energy/Oil_Reserves.csv")
colnames(Oil_Reserve) <- as.character(unlist(Oil_Reserve[2,]))
Oil_Reserve <- Oil_Reserve[c(-1:-3),c(-2,-3,-9)]
colnames(Oil_Reserve)[colnames(Oil_Reserve) == "" ] <- 'Region'
Oil_Reserve <- Oil_Reserve %>% gather(Year,Reserves_in_BillionBarrels,
'2012','2013','2014','2015','2016',-Region)

```

```

Oil_Reserve1 <- split( Oil_Reserve , f = Oil_Reserve$Year)
a <- Oil_Reserve1[1]
a <- as.data.frame(a)
a$X2012.Reserves_in_BillionBarrels
= as.numeric(a$X2012.Reserves_in_BillionBarrels)
a$percentage = (a$X2012.Reserves_in_BillionBarrels
/ sum(a$X2012.Reserves_in_BillionBarrels)) * 100
colnames(a)
<- c("Region","Year","Reserves_in_billion_barrels","Percentage")

```

```

b <- Oil_Reserve1[2]
b <- as.data.frame(b)
b$X2013.Reserves_in_BillionBarrels
= as.numeric(b$X2013.Reserves_in_BillionBarrels)
b$percentage = (b$X2013.Reserves_in_BillionBarrels
/ sum(b$X2013.Reserves_in_BillionBarrels)) * 100
colnames(b) <-
c("Region", "Year", "Reserves_in_billion_barrels", "Percentage")

c <- Oil_Reserve1[3]
c <- as.data.frame(c)
c$X2014.Reserves_in_BillionBarrels
= as.numeric(c$X2014.Reserves_in_BillionBarrels)
c$percentage = (c$X2014.Reserves_in_BillionBarrels
/ sum(c$X2014.Reserves_in_BillionBarrels)) * 100
colnames(c) <-
c("Region", "Year", "Reserves_in_billion_barrels", "Percentage")

d <- Oil_Reserve1[4]
d <- as.data.frame(d)
d$X2015.Reserves_in_BillionBarrels
= as.numeric(d$X2015.Reserves_in_BillionBarrels)
d$percentage = (d$X2015.Reserves_in_BillionBarrels
/ sum(d$X2015.Reserves_in_BillionBarrels)) * 100
colnames(d) <-
c("Region", "Year", "Reserves_in_billion_barrels", "Percentage")

e <- Oil_Reserve1[5]
e <- as.data.frame(e)
e$X2016.Reserves_in_BillionBarrels
= as.numeric(e$X2016.Reserves_in_BillionBarrels)
e$percentage = (e$X2016.Reserves_in_BillionBarrels
/ sum(e$X2016.Reserves_in_BillionBarrels)) * 100
colnames(e) <-
c("Region", "Year", "Reserves_in_billion_barrels", "Percentage")

Oil_Reserve <- rbind(a,b,c,d,e)
write.csv(Oil_Reserve, file = "Oil_Reserve.csv", row.names = FALSE)
Oil_Reserve$Percentage <- format(round(Oil_Reserve$Percentage, 2),
nsmall = 2)

Gas_Reserve <- Gas_Reserve[order(Gas_Reserve$Region),]
Oil_Reserve <- Oil_Reserve[order(Oil_Reserve$Region),]
Oil_Reserve <- Oil_Reserve[c(1,2,4)]
Gas_Reserve$Fuel_Type <- rep("Gas", nrow(Gas_Reserve))
Oil_Reserve$Fuel_Type <- rep("Oil", nrow(Oil_Reserve))
colnames(Oil_Reserve)
<- c("Region", "Year", "Reserves_in_Percentage", "Fuel_Type")

```

```

middle <- sqldf("Select Region,Year,sum(Reserves_in_Percentage)
as Reserves_in_Percentage,
Fuel_Type from Gas_Reserve where Region='Africa' or Region
='Middle East' group by Year")
middle$Region <- gsub("Middle East",middle$Region,"Middle East and Africa")
Gas_Reserve <- sqldf("Select * from Gas_Reserve where
Region!='Africa' and Region!='Middle East'")
Gas_Reserve <- rbind(Gas_Reserve,middle)

middle <- sqldf("Select Region,Year,sum(Reserves_in_Percentage)
as Reserves_in_Percentage,
Fuel_Type from Oil_Reserve where Region='Africa' or Region
='Middle East' group by Year")
middle$Region <- gsub("Middle East",middle$Region,"Middle East and Africa")
Oil_Reserve <- sqldf("Select * from Oil_Reserve where
Region!='Africa' and Region!='Middle East'")
Oil_Reserve <- rbind(Oil_Reserve,middle)

Oil_Reserve[order(Oil_Reserve[,1]),]
Gas_Reserve[order(Gas_Reserve[,1]),]

Reserve <- rbind(Gas_Reserve,Oil_Reserve)
Reserve$Fuel_ID <- as.integer(factor(Reserve$Fuel_Type))
Reserve <- Reserve[order(Reserve$Region),]
Reserve$Region_ID <- rep(c(1,2,4,5,3),each=10)

```

```
write.csv(Reserve,file = "Reserve.csv",row.names = FALSE)
```

Data Cleaning and Transformation for Raw GDP Growth

```

library(stringr)
library(tesseract)
library(splitstackshape)

getwd()
setwd("C:/Users/Pavan Kumar/Desktop/DWBI/WorkingDirectory/")
eng <- tesseract("eng")
text <- ocr("https://www.wto.org/images/img_press/768tbl1_e.png",
engine = eng)
North_America <- as.data.frame(substr(text,854,892))
South_Central_America <- as.data.frame(substr(text,893,943))
Europe_CIS <- as.data.frame(substr(text,944,975))
Asia_Pacific <- as.data.frame(substr(text,976,1004))
Middle_East <- as.data.frame(substr(text,1005,1044))

colnames(North_America) <- c("xxy")
North_America <- cSplit(North_America, "xxy", "_")
North_America[1,1] <- paste(North_America$xxy_1,North_America$xxy_2)
North_America <- North_America[,c(1,3,4,5,6,7)]
colnames(North_America) <- c("Region","2012","2013","2014","2015","2016")

```

```

colnames(South_Central_America) <- c("xxy")
South_Central_America <- cSplit(South_Central_America,"xxy","_")
South_Central_America[1,1]
<- paste(South_Central_America$xyy_01,South_Central_America$xyy_02,

South_Central_America$xyy_03,South_Central_America$xyy_04)
South_Central_America <- South_Central_America[,c(1,5,6,7,8,10)]
colnames(South_Central_America)
<- c("Region","2012","2013","2014","2015","2016")


colnames(Europe_CIS) <- c("xxy")
Europe_CIS <- cSplit(Europe_CIS,"xxy","_")
Europe_CIS[1,1] <- paste(Europe_CIS$xyy_1,"and_CIS")
Europe_CIS <- Europe_CIS[,c(1:6)]
colnames(Europe_CIS) <- c("Region","2012","2013","2014","2015","2016")


colnames(Asia_Pacific) <- c("xxy")
Asia_Pacific <- cSplit(Asia_Pacific,"xxy","_")
Asia_Pacific[1,1] <- paste(Asia_Pacific$xyy_1,"_Pacific")
Asia_Pacific <- Asia_Pacific[,c(1:6)]
colnames(Asia_Pacific) <- c("Region","2012","2013","2014","2015","2016")


colnames(Middle_East) <- c("xxy")
Middle_East <- cSplit(Middle_East,"xxy","_")
Middle_East[1,1] <- "Middle_East"
Middle_East <- Middle_East[,c(1,4,5,6,7,8)]
colnames(Middle_East) <- c("Region","2012","2013","2014","2015","2016")
GDP_Growth <- rbind(North_America,South_Central_America,Europe_CIS,
Asia_Pacific,Middle_East)
colnames(GDP_Growth) <- c("Region","2012","2013","2014","2015","2016")
GDP_Growth$`2012` <- as.numeric(GDP_Growth$`2012`)
GDP_Growth$`2013` <- as.numeric(GDP_Growth$`2013`)
GDP_Growth$`2014` <- as.numeric(GDP_Growth$`2014`)
GDP_Growth$`2015` <- as.numeric(GDP_Growth$`2015`)
GDP_Growth$`2016` <- as.numeric(GDP_Growth$`2016`)
GDP_Growth[3,2] <- -0.2
GDP_Growth[2,5] <- -1.0
GDP_Growth[2,6] <- -1.7
GDP_Growth <- GDP_Growth %>%gather
(Year,Growth_in_Percentage,`2012`,`2013`,`2014`,`2015`,`2016`)

GDP_Growth <- GDP_Growth[order(GDP_Growth[,1]),]
GDP_Growth$Region_ID <- rep(c(4,5,2,1,3),each=5)
GDP_Growth$Region <- gsub(pattern = "Middle_East",
GDP_Growth$Region,replacement = "Middle_East_and_Africa")
write.csv(GDP_Growth,file="GDP_Growth.csv",row.names = FALSE)

```

Data Cleaning and Transformation for Raw Population

```

getwd()
library(dplyr)
library(tidyr)
library(xml2)
library(rvest)
library(sqldf)
setwd("C:/Users/Pavan_Kumar/Desktop/DWBI/WorkingDirectory/")
Population <-
read.csv("C:/Users/Pavan_Kumar/Desktop/DWBI/Energy/Population.csv")
Population <- Population[c(-1:-3),c(-2:-56,-62)]
row.names(Population) <- c(1:nrow(Population))
colnames(Population) <- as.character(unlist(Population[1,]))
colnames(Population)[colnames(Population) == "1" ] <- 'Country'
Population <- Population[-1,]
Population <- Population[order(Population[,1]),]
row.names(Population) <- c(1:nrow(Population))
Population <- Population[c(3,9,12,20,28,37,44,45,46,56,66,82,87,110,111,
112,117,119,121,127,150,157,173,175,178,181,194,195,200,201,206,218,125
,222,233,237,238,244,249,250,251,252,255,257),]
row.names(Population) <- c(1:nrow(Population))
Population$'53' <- as.character(Population$'53')
Population[36,1] <- "Taiwan"
Population <- Population %>% gather(Year,Population,-'53')
colnames(Population) <- c("Country","Year","Population")
Population <- Population[c(2,1,3)]
Population$Region <-
ifelse(Population$Country == 'Canada',"North_America",
ifelse(Population$Country == 'United_States',"North_America",
ifelse(Population$Country == 'Argentina',"South_and_Central_America",
ifelse(Population$Country == 'Brazil',"South_and_Central_America",
ifelse(Population$Country == 'Chile',"South_and_Central_America",
ifelse(Population$Country == 'Colombia',"South_and_Central_America",
ifelse(Population$Country == 'Mexico',"South_and_Central_America",
ifelse(Population$Country == 'Venezuela',"South_and_Central_America",
ifelse(Population$Country == 'China',"Asia_Pacific",
ifelse(Population$Country == 'India',"Asia_Pacific",
ifelse(Population$Country == 'Indonesia',"Asia_Pacific",
ifelse(Population$Country == 'Japan',"Asia_Pacific",
ifelse(Population$Country == 'Malaysia',"Asia_Pacific",
ifelse(Population$Country == 'South_Korea',"Asia_Pacific",
ifelse(Population$Country == 'Taiwan',"Asia_Pacific",
ifelse(Population$Country == 'Thailand',"Asia_Pacific",
ifelse(Population$Country == 'Australia',"Asia_Pacific",
ifelse(Population$Country == 'New_Zealand',"Asia_Pacific",
ifelse(Population$Country == 'South_Africa',"Middle_East_and_Africa",
ifelse(Population$Country == 'Algeria',"Middle_East_and_Africa",
ifelse(Population$Country == 'Egypt',"Middle_East_and_Africa",
ifelse(Population$Country == 'Nigeria',"Middle_East_and_Africa",
ifelse(Population$Country == 'Iran',"Middle_East_and_Africa",
ifelse(Population$Country == 'Kuwait',"Middle_East_and_Africa",
ifelse(Population$Country == 'Saudi_Arabia',"Middle_East_and_Africa",

```

```

ifelse(Population$Country == 'United_Arab_Emirates',"Middle_East
and_Africa","Europe_and_CIS"))))))))))))))))))))))))))))
Population$Country_ID <- as.integer(factor(Population$Country))
Population$Region_ID <- as.integer(factor(Population$Region))
Population <- Population[c(1,6,4,5,2,3)]
write.csv(Population,file="Population.csv", row.names = FALSE)

```

SQL Queries for Raw,Fact table creation and Loading

\$Raw-Table-Creation

```

CREATE TABLE [Raw_Energy] (
    [Year] varchar(50),
    [Fuel_ID] numeric,
    [Fuel_Type] varchar(50),
    [Region_ID] numeric,
    [Region] varchar(50),
    [Country_ID] numeric,
    [Country] varchar(50),
    [Consumption] float,
    [Production] float,
    [Trade] float
)

CREATE TABLE [Raw_Reserve] (
    [Year] varchar(50),
    [Fuel_ID] numeric,
    [Fuel_Type] varchar(50),
    [Region_ID] numeric,
    [Region] varchar(50),
    [Reserves_in_Percentage] float
)

CREATE TABLE [Raw_Co2_Emission] (
    [Year] varchar(50),
    [Fuel_ID] numeric,
    [Fuel_Type] varchar(50),
    [Region_ID] numeric,
    [Region] varchar(50),
    [Country_ID] numeric,
    [Country] varchar(50),
    [Co2_Emission_in_mtCo2] float
)

CREATE TABLE [Raw_GDP_Growth] (
    [Region] varchar(50),
    [Year] varchar(50),

```

```
        [Growth_in_Percentage] float,  
        [Region_ID] numeric  
    )
```

```
CREATE TABLE [Raw_Population] (  
    [Year] varchar(50),  
    [Region_ID] varchar(max),  
    [Region] varchar(max),  
    [Country_ID] varchar(max),  
    [Country] varchar(max),  
    [Population] numeric  
)
```

#Dimension Table Creation

```
if object_id ('facttable','U') is not null  
drop table facttable ;  
go  
if object_id ('Dim_Location','U') is not null  
drop table Dim_Location ;  
go  
if object_id ('Dim_Fuel_type','U') is not null  
drop table Dim_Fuel_type ;  
go  
if object_id ('Dim_Time','U') is not null  
drop table Dim_Time ;  
go
```

```
create table Dim_Location (  
    Location_ID numeric IDENTITY primary key not null,  
    Region varchar(max),  
    Country varchar(max))  
  
insert into Dim_Location(  
    Region,Country)  
select distinct Region,Country from Raw_Co2_Emission order by Country
```

```
create table Dim_Fuel_type(  
    Fuel_ID numeric IDENTITY primary key not null,  
    Fuel_Type varchar(max)  
)  
Insert into Dim_Fuel_type(  
    Fuel_Type)  
select distinct Fuel_Type from Raw_Energy
```

```
create table Dim_Time(  
    Year_ID numeric IDENTITY primary key not null,  
    Year varchar(max))  
insert into Dim_Time(  
    Year)
```

```
Year)
Select distinct Year from Raw_Energy
```

Fact Table Loading

```
create table FactTable (
Fuel_ID numeric not null,
Location_ID numeric not null,
Year_ID numeric not null,
Co2_Emission_in_mtCo2 float,
Consumption float,
Production float,
Trade float,
Reserves_in_Percentage float,
GDP_Growth_in_Percentage float,
[Population] float)
```

```
Alter table FactTable add
constraint Fuel_fk Foreign key(Fuel_ID) REFERENCES Dim_Fuel_Type(Fuel_ID),
constraint Location_fk Foreign key(Location_ID)
REFERENCES Dim_Location(Location_ID),
constraint Time_fk Foreign key(Year_ID) REFERENCES Dim_Time(Year_ID)
```

```
insert into FactTable(
Fuel_ID ,
Location_ID ,
Year_ID,
Co2_Emission_in_mtCo2,
Consumption,
Production,
Trade,
Reserves_in_Percentage,
GDP_Growth_in_Percentage,
[Population]
)
```

```
select distinct a.Fuel_ID,
dl.Location_ID,
dt.Year_ID,
a.Co2_Emission_in_mtCo2 ,
b.Consumption ,
b.Production ,
b.Trade,
c.Reserves_in_Percentage,
d.Growth_in_Percentage,
e.[Population]
from Raw_Co2_Emission a join Raw_Energy b on
a.year=b.year and a.Region_ID =
```



```
b.Region_ID and a.Country_ID=b.Country_ID and a.Fuel_ID = b.Fuel_ID
join Raw_Reserve c on a.Region_ID =
c.Region_ID and a.Year = c.Year and a.Fuel_ID = c.Fuel_ID
join Raw_Population e on a.Year = e.Year and a.Country_ID = e.Country_ID
join Raw_GDP_Growth d on c.Year = d.Year and c.Region_ID = d.Region_ID
join Dim_Location dl on dl.Country = a.Country
join dim_time dt on dt.year = a.year
```