Abstract

This Project is all about extracting data from data warehouse which visualizing it and analysing it and generating Business Intelligence help to FMCG Companies. The project is automated in such way that it analyse past key relationship between agriculture data with government data and FMCG market. Project will also show about method and medium used to deliver this project. This project will provide clear view of Agriculture growing Impact on Indonesia economy in various ways, different forms of dataset analysed for purpose of generating another form of modulated data with respect to how Agriculture is defining process to run a Country, incorporating it with equipment’s and man force for FMCG industries to generate a desired balanced revenue. Since full automation used in project the desired result can be stored in digital form to help government and industries realize what the trend they need to follow for successful business setup. Project will create huge awareness among people to optimize use of tools since database will be in granulated form looking each aspect of agriculture, which in turn will also help indirectly government to reorder process for people benefits. If this project implementation, results are adopted it can coherent the relationship of large consumer base related to agro food with FMCG companies which can make agriculture as centre of force which can hold all industry.

1 Introduction

Agriculture Industry is evergreen industries which has indirectly huge impact on every sector, every people because it provide daily necessities. So there is huge increase in data because of complex relationship it has with every sector. Complex data can be structured if re-ordered in proper format using automated tools. For this data warehouse technique is used which can store large complex data in simple ordered form by restructuring data related to every recorded dimension which effecting data. The structured data prepared will be having past data and creating idea for future step which will be convenient.

Data warehouse project is designed in such way that will reduce time, cost, which indirectly will increase efficiency, **because it will serve as experienced machine, which will provide data in intelligence form** for which we need less effort to realize process required for implementation. This project will be beneficial for companies and government by letting them know where, & how to setup business .Data created by it is non-volatile which can be served in digital way across multiple server system. Decision making capacity improves because data is in dynamic visualized form where different relationship can be build and compared. It is an upgraded version of static database which provide structured information fast and through reliable access. Data Warehouse is process oriented, concatenated, invariant, Non-Volatile Platform

This report will bestow reliable architectonics used to build data warehouse and will show implementation approach followed for this data warehouse project.

Report consists of following parameters given below:

1. Benchmarking and Architectonics used to build Data Warehouse. 2. Data Modelling – Structuring the design of schema by using raw datasets to create measurable value and adherence to dimension and online analysis of data by creating relationship between dimension in variant form. 3. Extract, Transform and Load(ETL) – knowledge about creation of ETL from scratch, incorporating with feasible modern technologies in ETL, programming ETL flow, describing benchmarking used for ETL.

4. Business Intelligence – Different generated Business Queries will be highlighted and explained in this project with tools used and how all datasets have been related to each other to automate the process of constructing, critical estimate of Business Queries using appropriate automated tools. For this project Ralph Kimball methodologies is used because it relates to the big scale project, here discussion will be about why Ralph Kimball approach is appropriate, for big business projects, user based, medium budget with drill down approach through building information transfer , different categorized data and Structuring it proper way so it can be understood by defined user.

Business queries which generated digitally for this project after fully automated Data Warehouse was served with operational data.

1. Do employment of specific region can be incorporated somewhere else than agriculture industry which will boost industrial production?

2. Fertilizer Industry growth with use of agriculture land which can lead to more productivity?

3. FMCG companies possible growth with possible increase of urban population?

2. DATA SOURCES

2.1 Source 1: STATISTA

The dataset downloaded from: <https://www.statista.com/statistics/603387/indonesia-urban-population/> is in structured form provides relation of information on urban population from year 2005-2018. These is related to the project in terms of business model to be set up for categorized population. This dataset obtained is static with yearly data which also gives idea about dimension to be build around measurable values and create relationship with other dataset. The dataset available from STATISA WAS PUBLISHED in Year 2018 which means dataset is updated which helps in defining proper data query with other dataset.

2.2 Source 2: OECD

The OECD dataset downloaded from: <https://data.oecd.org/agroutput/crop-production.htm/> is structured dataset provides columns of information on crop production from year 2009-2017. The OECD dataset downloaded from: <https://data.oecd.org/gdp/investment-by-asset.htm#indicator-chart> is structured dataset provides columns of information on cultivated asset from year 2009-2017. These is relevant to this project because this dataset addresses the business requirements listed in Section 1 by providing how can agriculture industry can help in industrial production.

2.3 Source 3: WORLD BANK DATA

The WORLD BANK DATA dataset downloaded from: <http://databank.worldbank.org/data/source/indonesia-database-for-policy-and-economic-research> is structured dataset provides columns of information on agriculture food export, gdp, agriculture land used, fertilizer production from year 2009-2017. These is relevant to this project because this dataset addresses the government attitude in focussing on its agriculture strength to help its economy.

2.4 Source 4: Investing.Com

BISI stock data is downloaded from stock chart from year from May 2007- Nov 2018 <https://www.investing.com/common/modules/js_instrument_chart/api/data.php?pair_id=101353&pair_id_for_news=101353&chart_type=area&pair_interval=month&candle_count=120&events=yes&volume_series=yes&period=max> UNVR stock data is downloaded from stock chart from year Jan 2004 – Nov 2018 <https://www.investing.com/common/modules/js_instrument_chart/api/data.php?pair_id=101622&pair_id_for_news=101622&chart_type=area&pair_interval=month&candle_count=120&events=yes&volume_series=yes&period=max> INDF stock data is downloaded from stock chart from year from Jan 2008- Nov 2018 <https://www.investing.com/common/modules/js_instrument_chart/api/data.php?pair_id=101443&pair_id_for_news=101443&chart_type=area&pair_interval=month&candle_count=120&events=yes&volume_series=yes&period=max> this all dataset is unstructured dataset provides columns of information of stock value of FMCG stock listed company in Indonesia from year 2009-2017. These is relevant to this project because this dataset addresses how FMCG companies are performing in Indonesia.

|  |  |  |
| --- | --- | --- |
| Source | Type | Brief Summary |
| STATISTA | STRUCTURED | Used because it was mandatory to get find one data set from here and this dataset connect to all my dataset with year and country same in it. |
| OECD | STRUCTURED | Have Suitable form of data in excel of all country which was cleaned to get details about specific country. |
| WORLD BANK DATA | STRUCTURED | The data obtained from here have micro level information  about each aspect needed for my project. |
| INVESTING.COM | UNSTRUCTURED | This data obtained because it is part of relationship which is required to connect my one dataset to another and make reliable analysing. |

Table 1: Summary of sources of data used in the project

3. Related Works

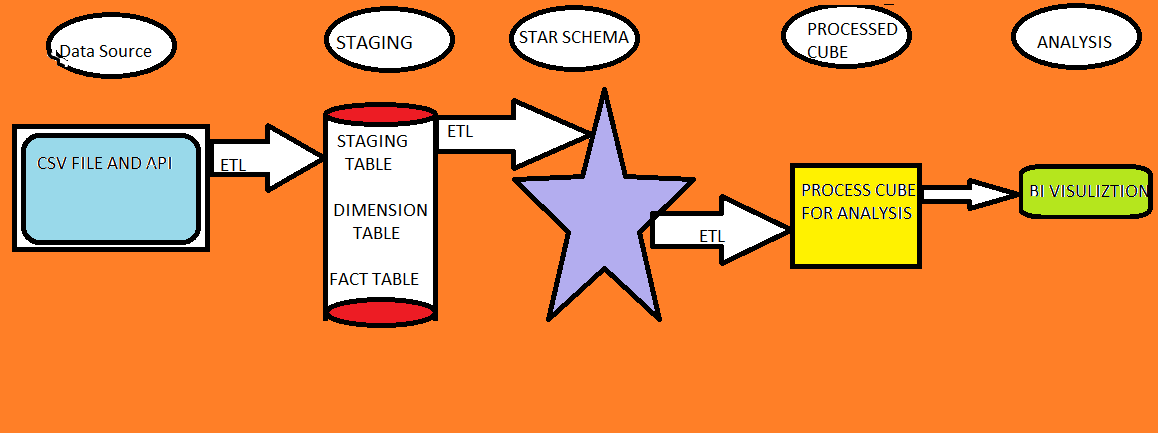
*Sustainable agriculture in Indonesia: Facts and challenges to keep growing in harmony with environment M. Faiz Syuaib* has been the journal which had previously worked on somewhat related to this topic. The dataset used in this literature review preparation is agriculture land measurement, crop production, fertilizer usage and urban population demographic to show the rapid pace of agricultural development in the last four decades as well as the commercialization, industrialization and urbanization which has led to significant changes in agricultural production systems. In some regions up gradation of agricultural technologies has increased production to keep pace with the population growth in urban areas.

This literature review is providing support to my project that if listed factors in dataset is worked properly it can benefit and create a new scope to our business requirement by developing suitable business intelligence. Agriculture has never been priority of any countries government because they don’t consider it’s a driving force for its economy, but consider just as a part of system which we need in our daily lifestyle. There has been always attempt to boost industrial production on verge of declining agriculture production by replacing agriculture land with industrial areas. Many people who work as farmer have lost their bread because of government policies which is making economic corridor for industrial companies. Due to which there is huge migration from rural areas to urban areas which has created a huge imbalance in urban development model. This whole scenario turns into vicious circle which can has long term effect in future by degrading life in urban areas.

The project can give results how can agriculture development and providing up gradation connection with industrialization to produce FMCG goods can create a proper balance of socio-economic development in country, by helping government prioritising its duties to help farmers of its country which can never be replaced. Implementing records data analytics in agriculture can boost its productivity by creating awareness among people in these occupation by making realizing them future scope.

4. DATA MODELLING

So, far in this project I have used bottom up Approach of Ralph Kimball, according to Kimball Approach we have to start with Star schema when we have fact tables and some dimension tables, fact will give you all measures whereas dimensions contains information, dimensions contains a primary which is used by foreign key in fact table.



Four Step of Data Modelling used in this project:

4.1 Define Business Process: - Business process used for this project is to provide efficiency in agriculture production and what impact it will have on industrial production by relationship with other factors. Finding stock price of FMCG company if going up, how production is going high, how employment is going high, according change the working style and improve in lacking area. Main motive of this analysis is to analyse agriculture data on yearly basis comparing it with gdp, export of food, employment, stock prices improve as required.

Today Agriculture Sector required to be on active in terms of business relationship and satisfaction of customer as whole. So, it is required to be efficient in terms of providing quantitative as well qualitative as whole and analysing the services given on time using data available from various countries on their site. So, I tried to provide a general predictive Analysis Engine which can be used by all countries in general to analyse the fact on yearly basis.

4.2 Defining grain of the Data Warehouse: Analysis will be done on yearly basis, so it will provide us information on yearly basis and analysis can be done on granular level and take effective measure instantly.

4.3 Create Dimensions: Identify the attributes from Data Tables and create separate dimension table for each of them.

Will try to provide information about what and where about business process, in this project we are using two-dimension tables (Year, Country). Year Dimension will be a Role-Playing Dimension used for many year.

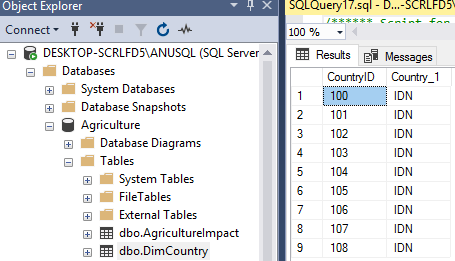


Figure 1: Dimension table of Country

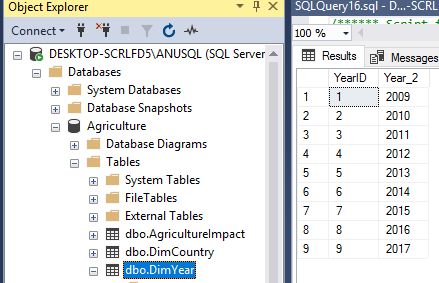


Figure 2: Dimension table of Year

Primary Key Used by dimension table to uniquely identify the row in the table. YearId, CountryID are primary keys for dimension tables shown in above tables.

4.4 Creating Fact table: After creating dimension create a fact table with all measures in it, these measures further be used for purpose of analysis.

Fact Table This table store all processed granulated measurement of business process, in this project review our analysis will be on yearly basis.

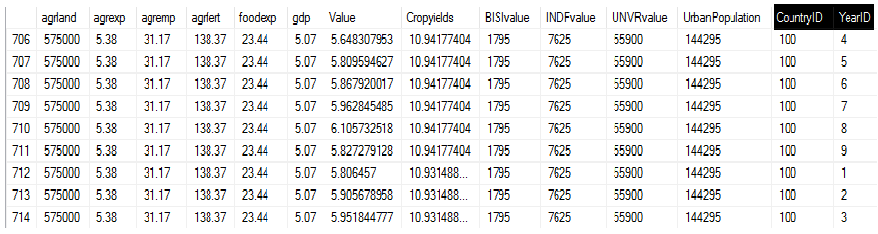
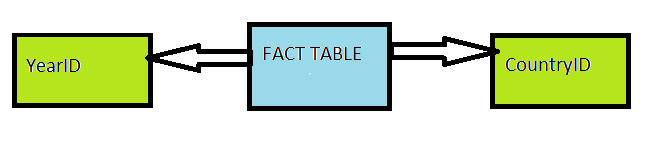


Figure 3: Fact Table

Foreign Key is a Column in Fact Table to reference another to create a Join between Fact and Dimension Table. In our project foreign key is CountryID, YearID.

4.5 Star Schema Design

Star Schema has a single table for each dimension, each table contains all attributes for that dimension, particularly a demoralized form. Figure 5: Raw Star Schema representation.

Star Join in Dimensional modelling is used to join both fact and Dimension tables, in start join facts are contained in fact table like agriculture land area, cultivated asset, stock value etc and information in Dimension table like Year dimension, Country dimension. Star join use both fact and dimension tables to build relationship and answer any query related to dimension and facts and up to most granular level of join.

5. Logical Data Map

Table 2: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 12

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | Column | Destination | Type | Transformation |
| 3 | CountryID | DimCountry | Dimension | Converted into Integer |
| 3 | agrland | AgricultureImpact | Fact | Converted into Integer |
| 3 | agrexp | AgricultureImpact | Fact | Converted into Percentage Form |
| 3 | agremp | AgricultureImpact | Fact | Converted into Percentage Form |
| 3 | agrfert | AgricultureImpact | Fact | Converted into Percentage Form |
| 3 | foodexp | AgricultureImpact | Fact | Converted into Percentage Form |
| 3 | gdp | AgricultureImpact | Fact | Converted into Percentage Form |
| 2 | YearId | DimYear | Dimension | Converted into Integer |
| 2 | Value | AgricultureImpact | Fact | Converted into Percentage Form |
| 2 | Cropyields | AgricultureImpact | Fact | Converted into Integer |
| 4 | BISIvalue | AgricultureImpact | Fact | Kept as it in currency form |
| 4 | INDFvalue | AgricultureImpact | Fact | Kept as it in currency form |
| 4 | UNVRvalue | AgricultureImpact | Fact | Kept as it in currency form |
| 1 | UrbanPopulation | AgricultureImpact | Fact | Kept as it in million form |

6. Overview of Extract Transform and Load

ETL is process of Extracting Data from raw sources, transforming it into measurable form and then loading it into Data Warehouse.

6.1 Extracting Data

Sources which are used in Data Warehouse, Sources can be any type Structured, Semi structured, Unstructured. In this project different sources of Data being used like Structured Data (CSV Files) downloaded from Websites, some data was scraped from website and some unstructured data is extracted from stock line graph using R Code which then converted to csv.

Two Types of Extractions methods used: 6.1.1 Logical: Full Extraction is extraction of data one time and no keys required in this extraction. Incremental Extraction is used when only changed data being extracted.

6.1.2 Physical: Online Extraction is done directly from source. Offline Extraction is done from Flat File, Dump File. In this project online Extraction done through R Code offline Extraction of Datasets Done using R Code.

• R Code used: given in Appendices I

• Data extracted is stored in SQL Server 2017 Database – Dimension Database for Dimensions and Facts, Stage Dimension is used as staging Area.

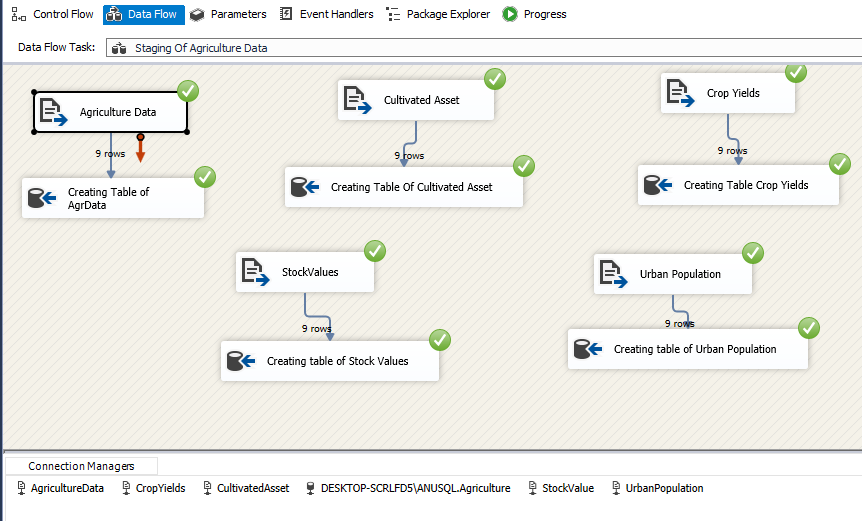


Figure 6: Screenshots of extractions used in this project.

We can see that flat file is loaded through connection manager in flat file source and then through SSIS pipeline into OLEDB Destination to create table in database of SQL server 2017 which is connected to SSIS through service analysis.

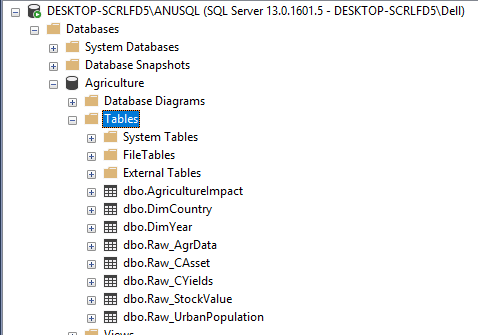


Figure 7: Screenshots of Showing Table Creation automatically after SSIS Execution.

6.2 Transformation: It is one of most difficult in terms of processing time, here we do simple data conversion to complex data aggregation, merging, editing. Number of transformations techniques are present in SQL Server Integration Services, but in these project minimum transformations is used as required for the project. Some Transformation done before staging Area and some before loading Data into dimensions and fact table. Different Transformations used are:

• Multistage Data transfer.

• Pipelined Transfer.

• Create Table using SQL.

• Use of Truncate, Lookup

• Multistage insert of Data.

Truncate is used by using SQL Command so that no values of table is repeated it contains all tables. It is used before staging.

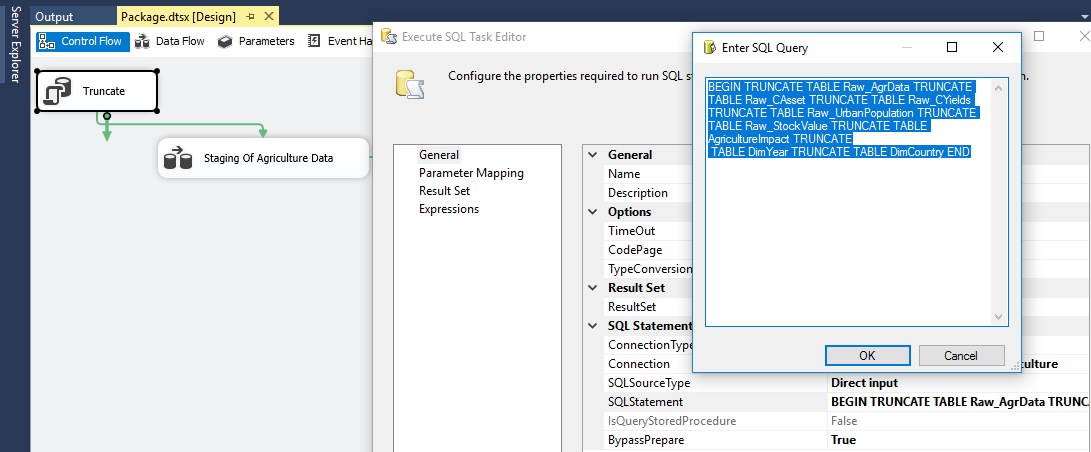


Figure 8: Screenshots Showing Truncation using SQL Command.

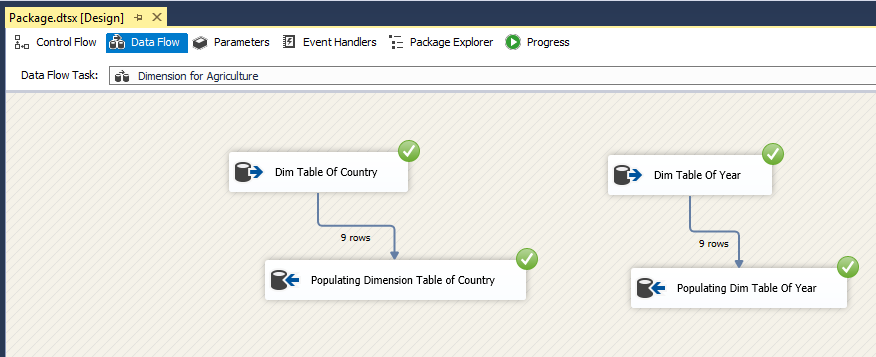
6.3 Loading: Loading Data into Dimensions and Fact: • Using SQL, used SQL code to Populate YearDimension and CountryDimension (Code Taken online) • Using pipelined multistage insert of Data.

Figure 9: Screenshots Showing Creation and Population of Dimension table using SQL Command.

• Using SQL, used SQL code to Populate Fact Table AgricultureImpact (Code Taken online) • Using pipelined multistage insert of Data. • Using lookup so dimension can be matched and there can be connection between them using primary key. •Foreign key is given between fact table and dimension so proper dimension can be created.

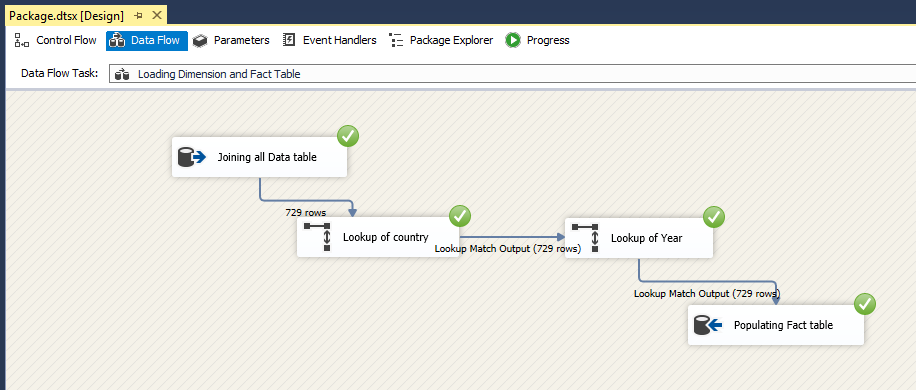


Figure 10: Screenshots Showing Creation and Population of Fact table using SQL Command and joining dimension to fact table by sing lookup.

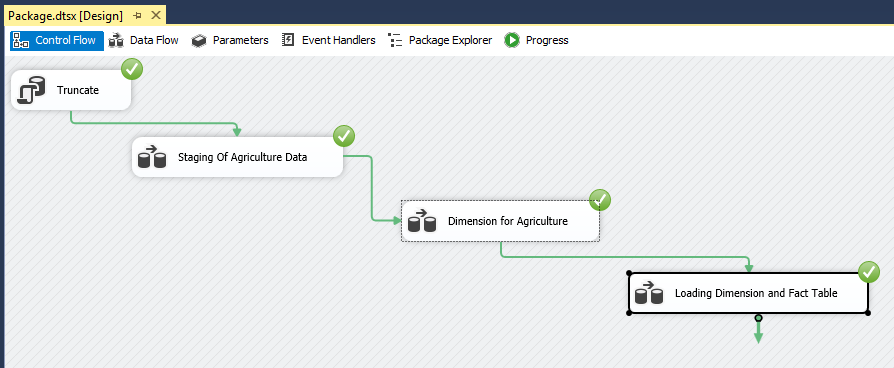


Figure 11: Screenshots Showing Creation and Population of all table and there flow from staging to Fact table.

•AgricultureImpact is Fact table which is connected to dimension year and dimension country by primary key connection from dimension key to fact table.

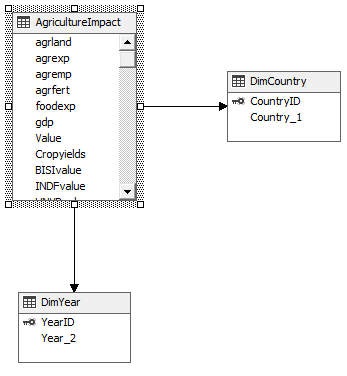


Figure 12: Star Schema Generation with primary key connected to fact table.

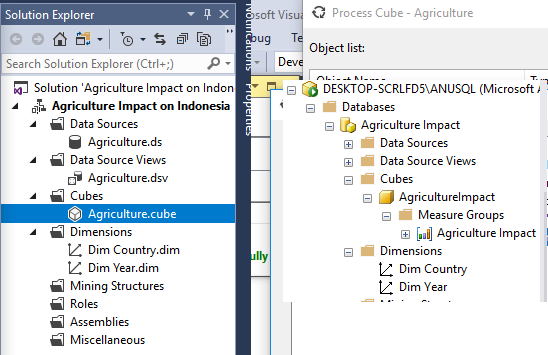
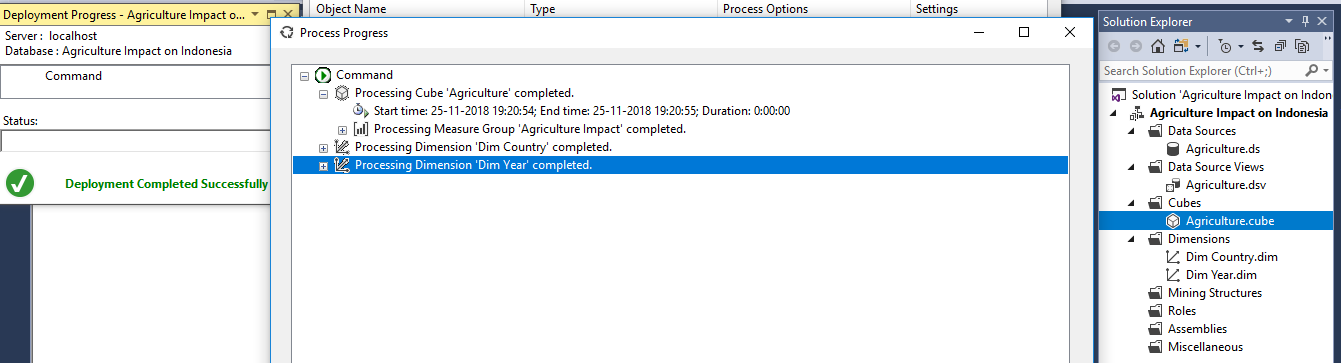


Figure 13: View of Cube in SQL Server Analysis Service and Analysis Server in SQL Server Management Studio

Figure 14: Successful deployment of Cube.

8. Business Intelligence

This project is designed to deliver proper ideas to government and FMCG companies where they should look for opportunity for their growth Cube was placed in essential column to get desired result and find answer to different Business queries, that’s why in the project I’ll be representing three main Business Queries which I found most impactful and provide insight in terms of choosing agriculture as base for answering queries on basis of yearly and historical data.

8.1 Business Query 1: Do less people employment in agriculture means more productivity?

Food export, crop yields, Gdp of Indonesia is not affected by people employment in agriculture so government can transform this workforce in industrial areas to gain maximum result in other areas and there might be reason that government is automating the field of agriculture production which shows less people employment in agriculture sector which in future create less dependency and continuous productivity.

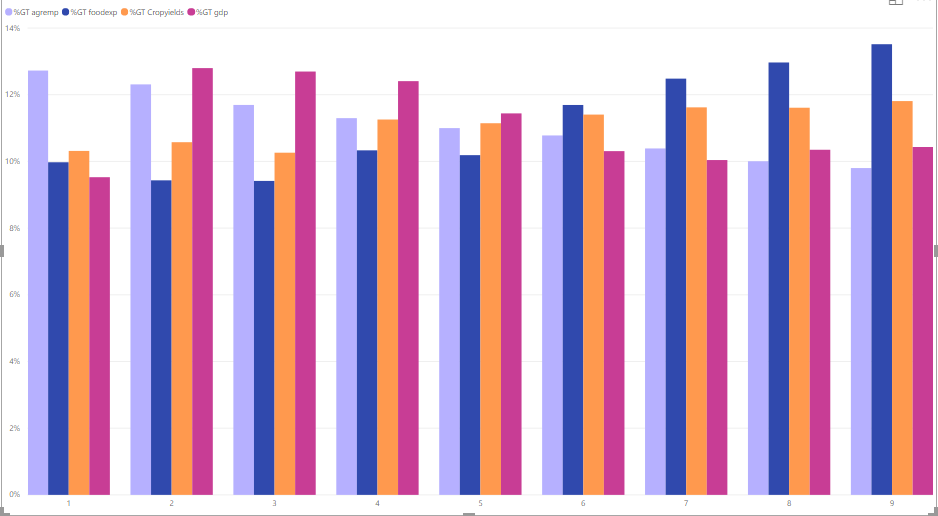




Figure 15: Screenshot results Of BI query1 from Power BI.

8.2 Business Query 2: Do fertilizer companies need to setup businesses in Indonesia?

Government of Indonesia is seriously considering taking agriculture as major field for its economic growth, since agriculture land usage in Indonesia has increased with time and use of fertilizer also shows that there is seriousness about productivity which can help to get more quantitative and qualitative production, since urban population is also increasing so there might be more requirement in coming years. Fertilizer Industry can benefit from it by getting great consumer base.

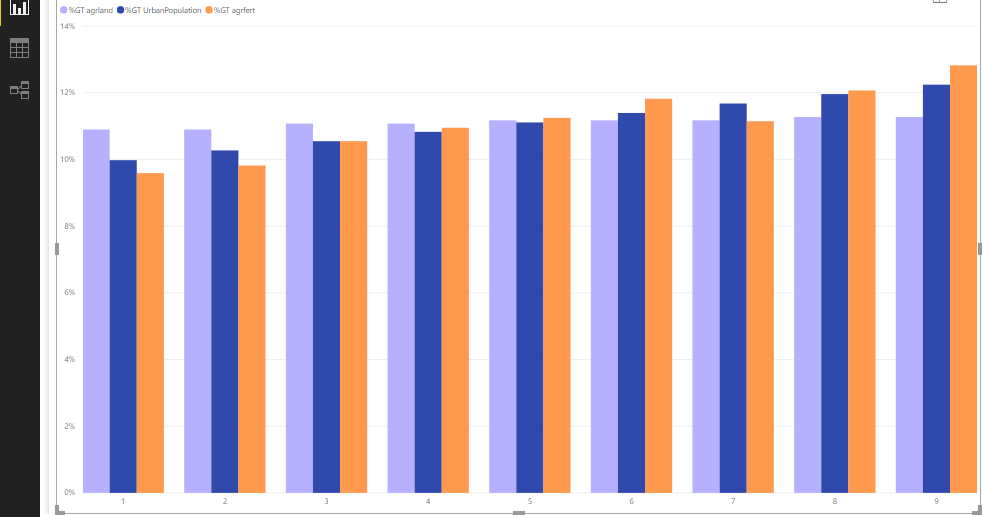




Figure 16: Screenshot results Of BI query2 from Power BI

8.3 Business Query 3: Do FMCG companies need to setup companies in Indonesia?

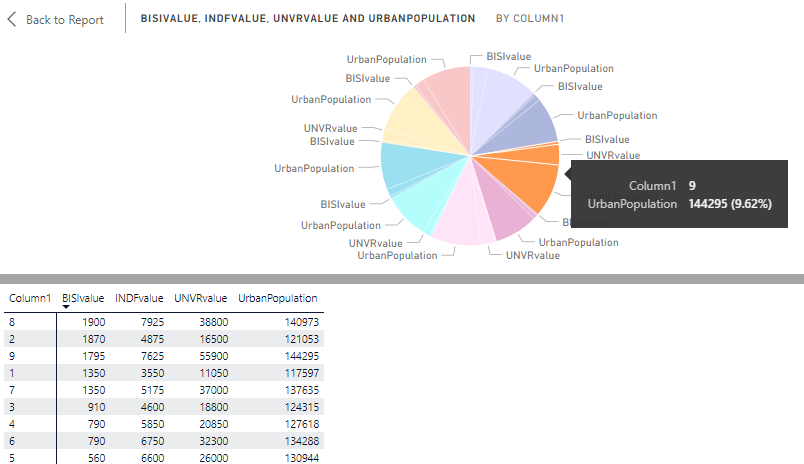
Urban Population in Indonesia has increased so has demand for food also increased with time productivity which can help FMCG companies to get great consumer base in Indonesia , because there has been rise in stock values of existing FMCG companies in Indonesia in recent years.

Figure 17: Screenshot results Of BI query3 from Power BI

8. Conclusion and Future work:

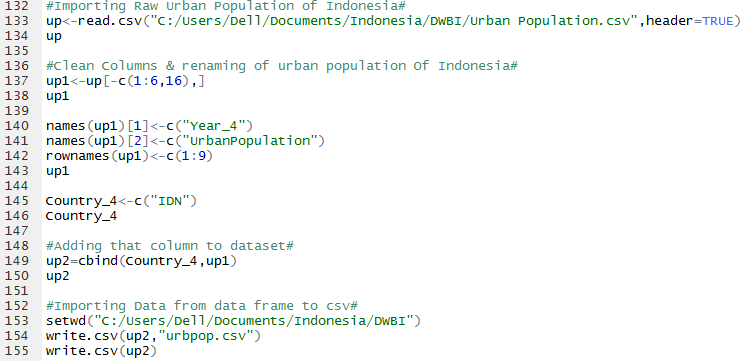
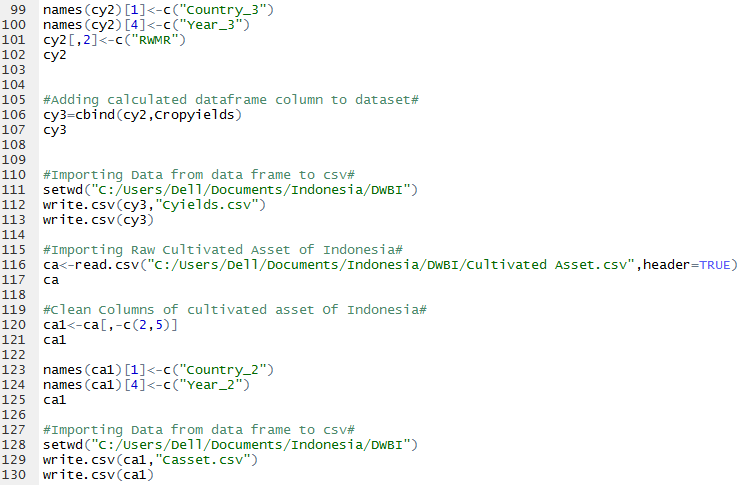
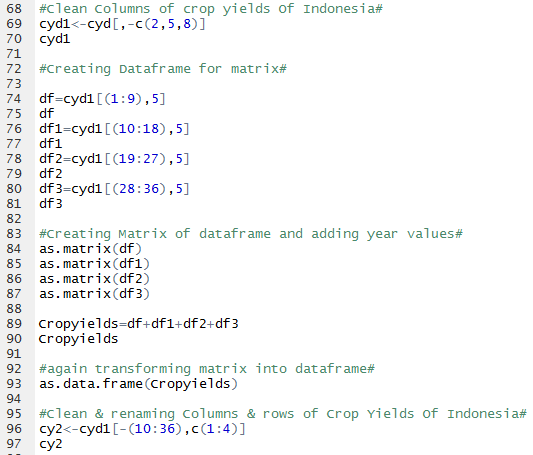
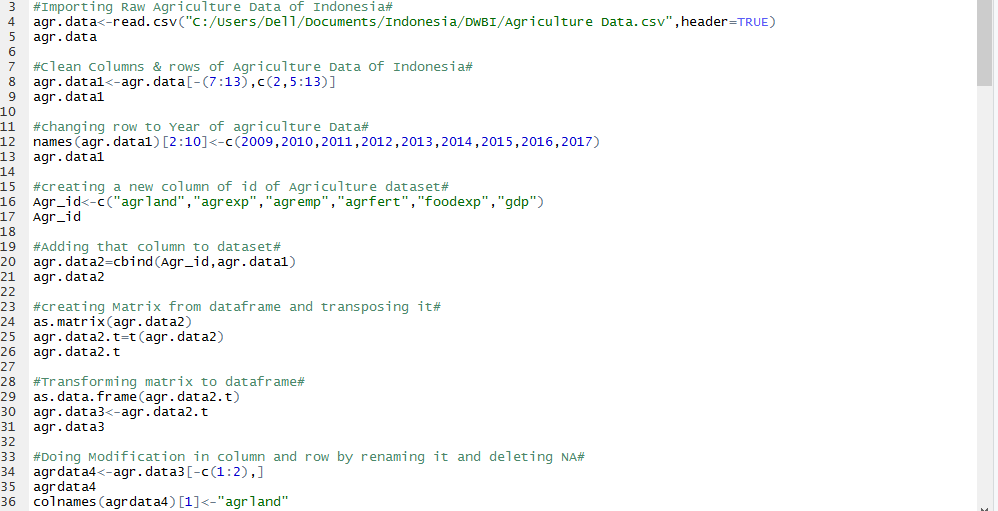
This project can be further modified in improved way by including rural areas, industrial companies producing agriculture equipment in project, but there is no data available from government, only two dimension are included in these project because complexity of agriculture can only be understand with respect to country and year. Running R Code and connecting it with SSIS to check monthly analysis of FMCG in particular city. We can also further automated process to find review analysis score for all food deliver app data for most influential cities in terms of food and using this monthly analysis Engine with comparing it with historical Data.

References

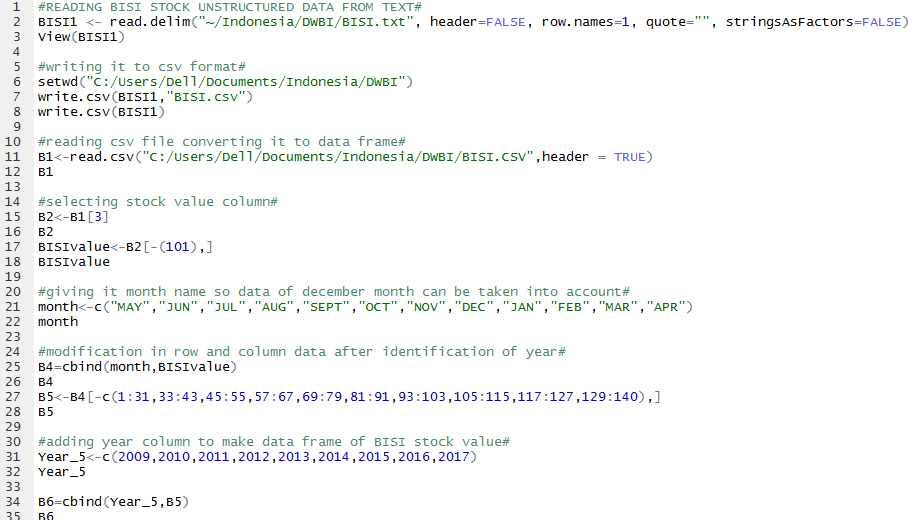
* *Sustainable agriculture in Indonesia: Facts and challenges to keep growing in harmony with environment M. Faiz Syuaibe*
* *Youtube*
* *Stackover flow*

Appendix

* R Code



Extracting stock data using R



Sql Query Code

