Data Warehousing and Business Intelligence Project

on

Impact of Agriculture On Indonesia

Anurag Abhay Singh
x18104053

MSc Data Analytics – 2018/19

Submitted to: Dr.Simon Caton

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

Student Name:	Anurag Abhay Singh
Student ID:	x18104053
Programme:	MSc Data Analytics
Year:	2018/19
Module:	Data Warehousing and Business Intelligence
Lecturer:	Dr.Simon Caton
Submission Due Date:	26/11/2018
Project Title:	Impact of Agriculture On Indonesia

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 28, 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 LaTeX 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

Impact of Agriculture On Indonesia

Anurag Abhay Singh
x18104053

November 28, 2018

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

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.

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

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

(Req-3) 3. FMCG companies possible growth with possible increase of urban population?

Source	Type	Brief Summary
STATISTA	STRUCTURED	Mandatory to get one data set from this site and one which links to all data set by having common attribute in them.The Statista dataset sets is presented in form of country and urban population
OECD	STRUCTURED	Data obtained in structured way which saves lot of time and have proper interlink to all other data tables. Data has variation in it which avoids redudancy.
BlOOMBERG	UNSTRUCTURED	Unstructured Data set was mandatory. This website was choosen because it creates uniqueness of my project by obtaining data rather than review which relates to my topic. The web site provided api key to acess data.
WORLD BANK DATA	STRUCTURED	Dataset obtained contains relational level information level about each factor related to project.

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

2 Data Sources

2.1 Source 1: STATISTA

The dataset downloaded from `https://www.statista.com/statistics/603387/indonesia-urban-population/` which is available on STATISTA website 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 dataset downloaded from: `https://data.oecd.org/agroutput/crop-production.htm/`from OECD website is in structured form provides related information on crop production from year 2009-2017, which will define what changes taken place with crop production on other dataset which would be having a correlation term with agriculture. The data extracted by changing tear parameter and selecting crop column from website. Data are updated daily on OECD site so there is no doubt on authentication of this data.017.

The dataset downloaded from: `https://data.oecd.org/gdp/investment-by-asset.htm#indicator-chart`is again a structured dataset provides variable of information on asset grenerated from cultivation from year 2009-2017. These is related highly to this project because this dataset somewhat give clear idea about business requirements listed in Section 1 by showing monetary value year wise, which is important parameter for setting up any business and relates to agriculture industry in many ways. Data are updated daily on OECD site so there is no doubt on authentication of this data.

2.3 Source 3: WORLD BANK DATA

The dataset downloaded from: `http://databank.worldbank.org/data/source/indonesia` WORLD BANK DATA is also structured dataset and one of the best dataset provides time variant information of different agriculture sector food export, agriculture land usage, fertilizer production from year 2009-2017. These is highly structured data which clearly co relates to other dataset. This dataset give clear motive of Indonesia government for agriculture sector and shows what plan they can make for future in agriculture sector. World Bank Data is website which looks for every country data in granulated form, comparing all aspects related to data. Thats why this one of the best genuine data.

2.4 Source : Bloomberg.com

Bloomberg.com is company which provides stock data on requesting of api, The data used is from year 2009-2017 so there is no issue of dynamically changing data. The data from these website can be compared with any stock website which gives same result which was checked multitimes so there is no issue about reliability of this site. The data obtained from this site is regarding FMCG companies stock value for given year. The data set obtained from this site is unstructured which has to be cleared to get structured data. These is related to this project because this dataset addresses how FMCG companies are performing in Indonesia which will help us to generate business query for government.

BISI International Tbk is the multinational company in Indonesia which produces major agriculture products. PT (BISI) stock data is downloaded from stock chart from year from May 2007- Nov 2018 `https://www.bloomberg.com/markets2/api/history/BISI%3AIJ/PX_LAST?timeframe=5_YEAR&period=weekly&volumePeriod=weekly` this multinational company produces major agricultureproducts.

Unilever Indonesia Tbk PT is multinational FMCG company in Indonesia(UNVR) stock data is downloaded from stock chart from year Jan 2004 – Nov 2018 `https://www.bloomberg.com/markets2/api/history/UNVR%3AIJ/PX_LAST?timeframe=5_YEAR&period=weekly&volumePeriod=weekly`

Indofood Sukses Makmur Tbk PT(INDF) is company in food ndustry domain stock data is downloaded from stock chart from year from Jan 2008- Nov 2018 `https://www.bloomberg.com/markets2/api/history/INDF%3AIJ/PX_LAST?timeframe=5_YEAR&period=weekly&volumePeriod=weekly`

3 Related Work

"Sustainable agriculture in Indonesia: Facts and challenges to keep growing in harmony with environment", Cigrjournal.org, 2018. [Online]. Available: http://www.cigrjournal.org/index.php/Ejounral/article/view/3747 is the journal which cited work on Agriculture in Indonesia. The data given by this journal as show how diversification in agriculture production has impacted export of country. The report represent challenges faced by agriculture even being the driving force in economy of Indonesia. The challenges is same which this project is working on to improve productivity and give providing connection between farmer and FMCG companies.The literature review also show the concern I was having while designing database for my project the rise urban population and its demand for food. The paper cites about using moder agriculture techniques incorporated with eqipments to increase production. The datas set used in this literature review is from World Bank. Environment at a Glance, Indonesia. http://siteresources.worldbank.org/ from which one of of my dataset is acessed. The terms used in checking Indonesia Agriculture is measurement of land alloted to agriculture sector, crop production, demographic view of population, people engaged in agriculture field, crop production according to demographic view. It highlights exponential pace at which major development changes are taking place in agriculture in the recent four decades as well as the improved supply chain management, industrial production and urban development has led to notable changes in agricultural production systems.Review cites about what implmentation has been done by Indonesia to keep required pace with different consumer base companies.

This literature review is adding support to project idea that if certain critical works properly desired long term result can be obtained. Provided dataset should have all that factor which is important for its proper functionality. Literature review helps in stating and helping business queries related to employment ,crop production, consumer market. The whole idea about choosing this as literatutre review as base to define scope of my project for business and government by business intelligence is worked properly it can benefit and create a new scope to our business requirement by developing suitable business intelligence. Considering Agriculture sector as a priority never been a choice of any government because it requires lot of time and planning to implement the existing modern and current process in it which don't give desired result even implemented properly because some of critical factors are left out while designing a output system for it.Government is always following trend of developed countries by doing same always attempting to boost industrial production on verge of agriculture production which leads to it declination by giving agriculture land to corporate house and to match growing population demand by building transport ways. For many people who have domain of agriculture have to replace their likelihood by choosing different field for survival which again causes agriculture production decline. Migration of people from rural to urban area might happen due to which urban development structure might falls which may again lead or force government to design urban development plan for country. This what whole happens like circle in which every factor comes to same place irrespective what developments has been made. Once this vicious circle starts everyone gets affected by it. The project give relationship how can agriculture development and industrialization can go hand in hand to maintain proper socio-economic balance. If proper supply chain used it can improve productivity plus get new consumer base, which can people of there domain irreplaecable and and important in terms of contribution to economic. Incorporating data analytics techniques a new market can be formed where by connecting through by knowing which terms of relationship interlinks people of different chain. the project will creatte a huge awarness among government ,people of agriculture domain by realizing them the future scope of their domain.

4 Data Model

This project is designed by using Ralph Kimball approach taken refernce from https://www.computerweekly.com/tip/Inmon-or-Kimball-Which-approach-is-suitable-for-your-data-warehouse bottom up approach is used,according to Kimball Approach the data warehouse architecture starts by designing star schema a which has dimension and measurable value.

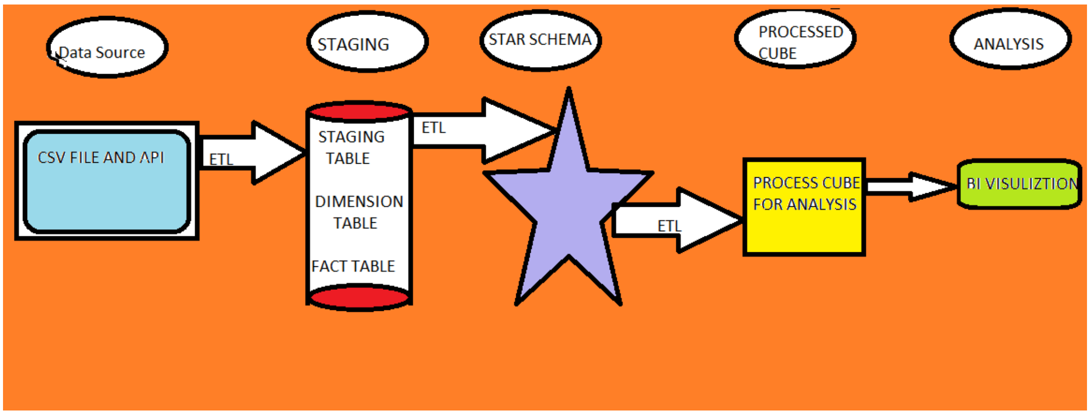


Figure 1: Data Warehouse Architecture by Ralph Kimball

Four Step of Data Modelling used in this project:

4.1 Defining a Business Process:-

Business process is term used to defining business model which is prepared by aggregating all business factors. Project is designed to connect efficiency in agriculture sector and correlated impact it generates on industrialization by having relationship with same factor in them. Correlating terms like price of stocklisted FMCG company with agriculture production or they are inversely propotional to each other which can help us in neglecting one of two factors or designing data warehouse including another term which co-relates with that other.Certainly if everything goes in proper way a new set of whole business intelligence is developed therefore the first step defining a business process act as a base to provide a proper data warehouse system. Business process developing defines which sector needs improvement and which terms can be added for its up gradation and improvement which makes that data warehouse a important tool to be noticed for future use. Data warehouse is important for business process because the data on which whole system works is dynamic so we need to just automate on existing data by addition and upgrading it and making a reliable connected data which can be shared on multi platform. Motive behind this business process is to generate business queries by analysing agro industries data on yearly basis comparing it with,FMCG stock, agriculture employment,government efforts and suggesting new techniques. Agriculture and industrialization need to be on active terms for building business relationship and creating a new consumer base as whole. So,the requirement is to be efficient in terms of providing measurable as well subjective product through supply chain management as whole and modifying the services through analysing it on time. This can be done by comparing data and building relationship between them. Project is clear in terms of providing digital form predictive analysis system which can be used by all countries in appropriate way to analyse the fact on yearly basis.

4.2 Defining grain of the Data Warehouse

Grain of data warehouse is common term which connects to parent database or form interlink between them. It is defined by dimension used to check measurable value in particular order. In project analysis is done yearwise and country wise so our grain is defined by heirarchies formed in year as month or date and in country by region and states. The more level of grain the complex the data warehouse and highly relational database we get.

4.3 Create Dimensions:

The most important factor for making data warehouse a dynamic and creative is creating Dimension. Attributes are identified which has relation between them and which connect to each other and form a whole new data table. Dimension table are created to define dimension which has stored data from different data table having heirarchies. Our project consist of two dimension table year and country. Dimension creates a new data and gives predictive visualization to end user. This sorts data and user has clear idea which dimension should be brought in use to develop new business queries. Primary Key The link to connection with measurable value is provided by primary key in dimension table. It is given to identify attributes of relational data which forms a connecting link to supply data and check data. CountryID and YearID are the two primary key used in the project.

4.4 Creating Fact table:

Fact table is data table which has all measurable values from different data table in sorted way and in unique form this fact table is connected to dimension table using surrogate key which forms interlink with dimension by comparing its measurable value with dimension available in dimension table generating new queries. Fact table measures on measurable values not character. measures in it, these measures further be used for purpose of analysis. fact table used in project compared with year and country.

4.5 Star Schema Design

Star Schema is formation of table which is has all values from dimension and fact table in by having interlink between them through keys of fact table and dimension table in normalized form. Star Join is used to create multidimensional modelling which is created by unification of both fact and dimension tables,Project consist of facts are which are related to measurable values like agriculture land area, cultivated asset, stock value etc and information in Dimension table like Year dimension, Country dimension. Star join use granulated level of join to connect both data available in fact and dimension tables to build relationship and generates business intelligence.

5 Logical Data Map

Table Containing Star Schema dimension table and fact table attibutes Details

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

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
1	urbpop	AgricultureImpact	Revenue	Fact	Primary Genre only
2	YearId	DimYear	Genre	Dimension	Converted to Integer
2	value	AgricultureImpact	Revenue	Fact	Primary Genre only
2	CropYields	AgricultureImpact	Percentage	Fact	Primary Genre only
3	CountryId	DimCountry	Genre	Dimension	Converted to Integer
3	agrland	AgricultureImpact	SqKM	Fact	Primary Genre only
3	agrfert	AgricultureImpact	Revenue	Fact	Primary Genre only
3	agremp	AgricultureImpact	Percentage	Fact	Primary Genre only
3	foodexp	AgricultureImpact	Revenue	Fact	Primary Genre only
3	agrexp	AgricultureImpact	Revenue	Fact	Primary Genre only
3	gdp	AgricultureImpact	Percentage	Fact	Primary Genre only
4	BISIvalue	AgricultureImpact	Revenue	Fact	Primary Genre only
4	INDFvalue	AgricultureImpact	Revenue	Fact	Primary Genre only
4	UNVRvalue	AgricultureImpact	Revenue	Fact	Primary Genre only

6 ETL Process

ETL is automated process which contains extraction tool to get data from any source and transferred through connected source to load tool to store same data in its database.The data loaded through ETL tool in database is in measurable form. The measurable form is used because read of measure is only possible in through ETL tool to give business intelligence.

6.1 Extracting Data

Sources can be of different form which is first read by ETL tools to assign various address bus to it, ex data which contains character type data is assigned 'char' address bus, and all measurable value are assigned various 'numeric','int','float' depending upon its value. Extraction is base of ETL because later that data needs to be related with various other data tables. ETL efficiency mostly depend upon its read operation command to sort all data properly for loading process. In project unstructured data was read and transferred to respective column of table.R programming code automation is used to get increase efficiency of reading operation and getting output in readable data format and which are used in Data Warehouse. Project contains data set of csv format which acts as source of data to ETL. Automated web scrapping is also possible through R code which act as extracting tool at initial phase for ETL. ETL is connected to remote database through service analysis link.Any changes is ETL directly automates the command which directly update the data in database storage system.

Two Types of Extractions command automatically happens in ETL only by reading data which makes it more efficient tool than any other tool 6.1.1 Logical Extraction: Full Extraction technique is used when static data need to be extracted which is not time variant. Incremental Extraction technique is process occurs when dynamic data need to be extracted which is time variant or data which is to be added with data in stored in database

6.1.2 Physical Extraction: Online data is extracted from data source through online extraction. Offline Extraction of data occurs from readable file format or saved file in computer by linking to it through connection.

In this project online and offline extraction both is used by R Code to feed data in database.

- R Code used: given in Appendix I

Connection manager is required to connect ETL with data source.Different data source have different connection manager to avoid confusion and redundancy of data.By accessing connection manager the created file can be viewed in database.This is reason ETL is highly efficient tool because it helps in creating highly optimized database at initial phase by sorting it in proper manner by giving it separate connection manager You can modify data from database or ETL and create a modified data. In project Flat File connection manager is used to extract flat file and loading in database. Multistaging process starts after completion of extraction process.

6.2 Transformation

It is one of most convenient in terms of processing data with respect to time,but require knowledge of all transformation tools.The important part which is beneficial for user of ETL is that the transformation operation can be changed,delete and copied also which saves lot of time by one way method of predefined code of database. Project has SQL server 2017 database for so transformation is done by using SQL query code. Different database have different query code which makes transformation tool user friendly for user of different server.The data extracted from is transformed to readable file and through SSIS pipeline it is transferred to Online Data base system. The file obtained from extraction tool is converted into data table in sorted way which is given by allocating each block to different data in rows and column of table. Transformation creates table where data easily get inserted by using column name given in flat file source. This reason makes ETL transformation tool more efficient because if inserted column name is not fed in query code table dosen't get created in database because it dosen't get required command from flat file source. To add on if user don't want to include certain column from flat file simply that column name is not inserted in query code which omits column data by default. Primary key connection is given by transformation tool. Transformation command used in project is Truncate,sort,lookup Truncate command was used to avoid repetition of value which ocurs when data table form relationship if they have same attribute. Sort command was used to sort dimension columns in proper order so that no values is repeated. Lookup is one of best command of ETL tool because it connects directly two dimension table to its data file where attribute matches avoiding query of codes. Transformation of data occurs at every stage of multistaging it connects extraction process to loading process through SSIS pipeline.

6.3 Loading

Loading is process of transferring transformed data in online database for actual use to generate business intelligence. Loading option can be recalled at each and every stage of multistaging. Loaded data cannot be transformed once stored only can be deleted to modify. SSIS pipeline is used to insert data into database. Transformed data are stored in form of table in database. Loading of data can be in bulk form if there are many values to be inserted. Incremental data loading is used to add new data in existing table by connecting it various dimension table and creating a new fact table connected to dimension table.Both option of loading has limitation that data cannot be changed. The data can only be change in extraction and transformation process. Loading is one way process. Loading option just show data in organized form which can help data base user to add incremental data if required for future use. Star Schema is generated after transformed data populated in dimension table and fact table. Loaded data is transferred through Service analysis to analyze different data while changing dimension.

|Cube Deployment Stage

The final stage includes making a multidimensional cube which is visualized by Data visualization tool in dynamic form. After reviewing star schema our deploying process starts which defines the shape of cube. Using cube is appropriate because it has all data stored in it. The cube is only made if there is proper link of database to ETL. Loading and unloading of data in database change the structure of cube because cube has properties of overwriting previous information because its multidimension dynamic modelling.

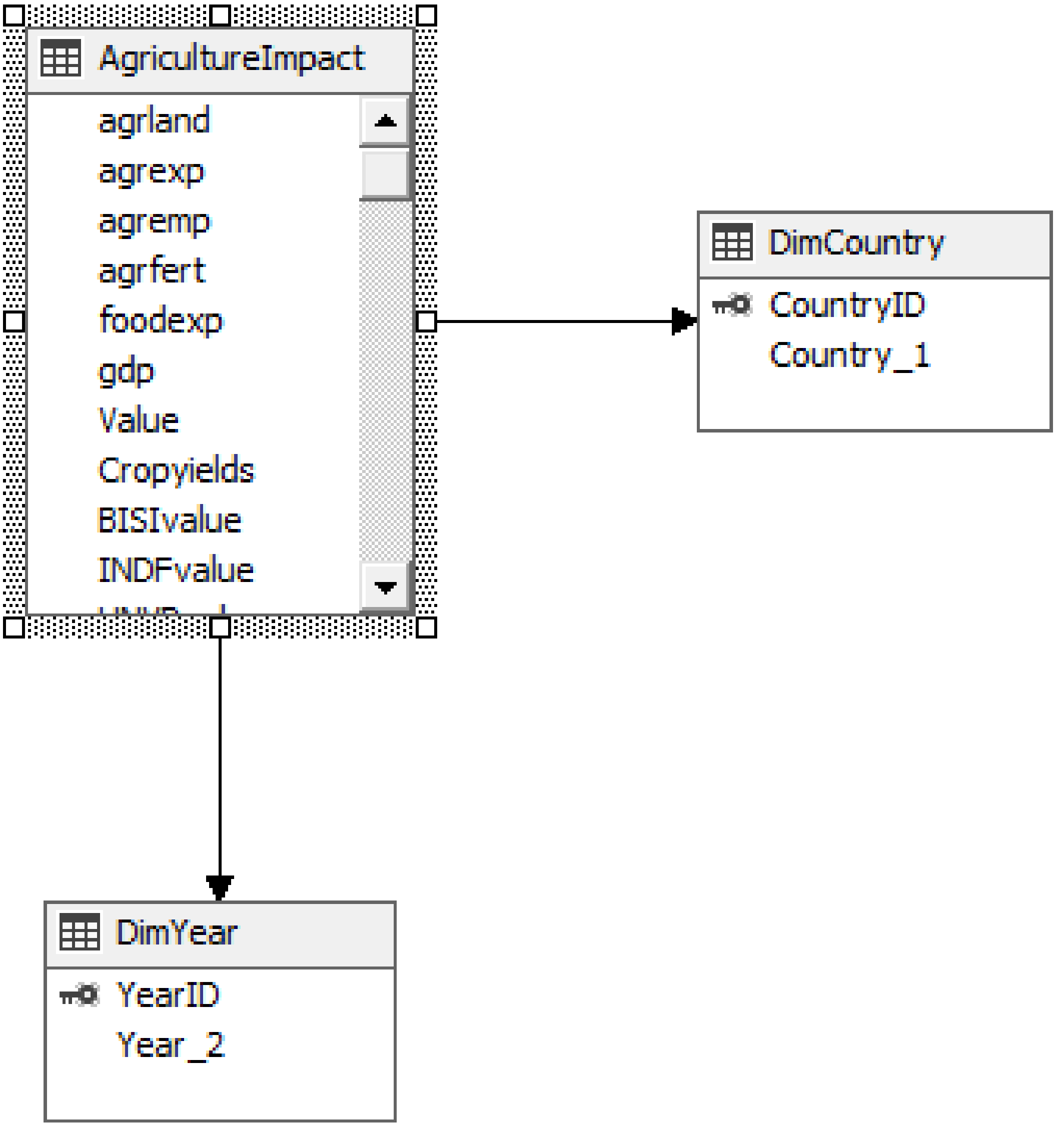


Figure 2: Star Schema Generation with primary key connected to fact table.

7 Application

Purpose of designing any automated system is not only to integrate it with existing system to reduce workloads but also to make improved decision making capacity. Our project has worked in different way it has given data visualization in such way that now we have highly reliable efficient decision making portable system with us. The data warehouse built in these project has delivered artificial business intelligence queries which will give proper direction in which companies can work mutually with others to obtain desired result. Application of these project will not be on agriculture sector but on every sector related to it, dependent on it. And since agriculture sector will never die only method of implementation change the flow of data will be continuous which again will generate new queries again and again we can easily work on our online database system to answer that query. The agriculture sector impact on Indonesia has delivered many business queries but its better to highlight only 3 business queries which will have deep impact on project.

7.1 BI Query 1: Agriculture sector employment is inversely proportional to agriculture productivity?

This Business query was generated by relationship between three different data source from OECD , WORLD BANK Data Statista .Food export, crop yields, Gdp are co related to each other which we can visualize but there is change in agriculture employment sector which bring us to predict that agriculture sector is not being considered as major source of income in Indonesia and the trend is toward automation which can maximize benefit of agro companies by using less manpower and the replace agriculture workforce can serve to other industry like of FMCG which is closely related to agriculture sector. Which in turn help government to give desired industrial production because less dependency will be on people.

Visualized Business Queries are illustrated in Bar graph form

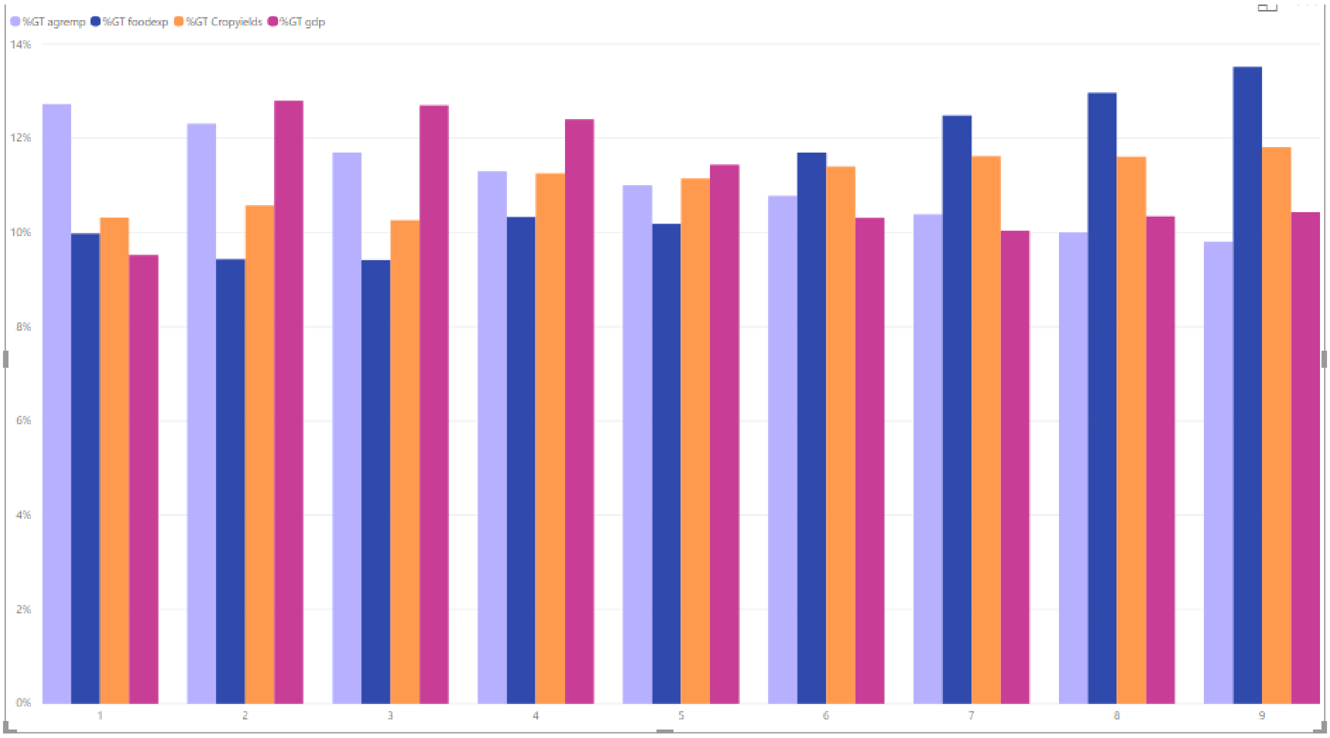


Figure 3: Results Of BI query1 from Power BI

7.2 BI Query 2:Business Query 2: Do fertilizer companies have scope in Indonesia?

This business query was generated when complex inter relationship was built from Data source World Data bank, OECD and Statista. The seriousness of these query can be understand by fertilizer companies which see Government policy of Indonesia working in there favour by increase usage of agriculture lands and increase use of agriculture by people and with that increase crop yields, so somewhat there is consumer base is slowly generating for fertilizer industry because gaining confidence of people is clearly seem to be happening in Indonesia. And people diversion from conventional method of farming to modern method which is stated in business query 1 can help fertilizer industry to maximize its production to achieve increased crop production and to create a large consumer base market of agro products.

Visualized Business Queries are illustrated in Bar graph form

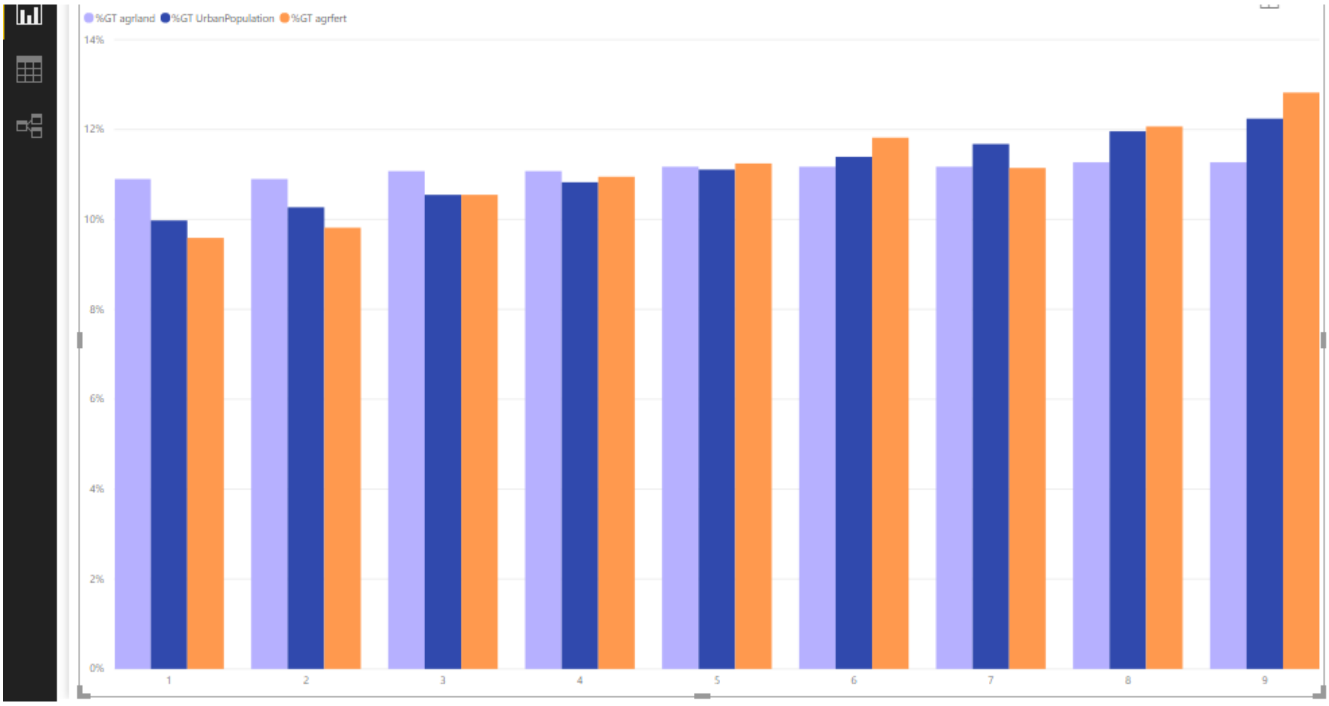
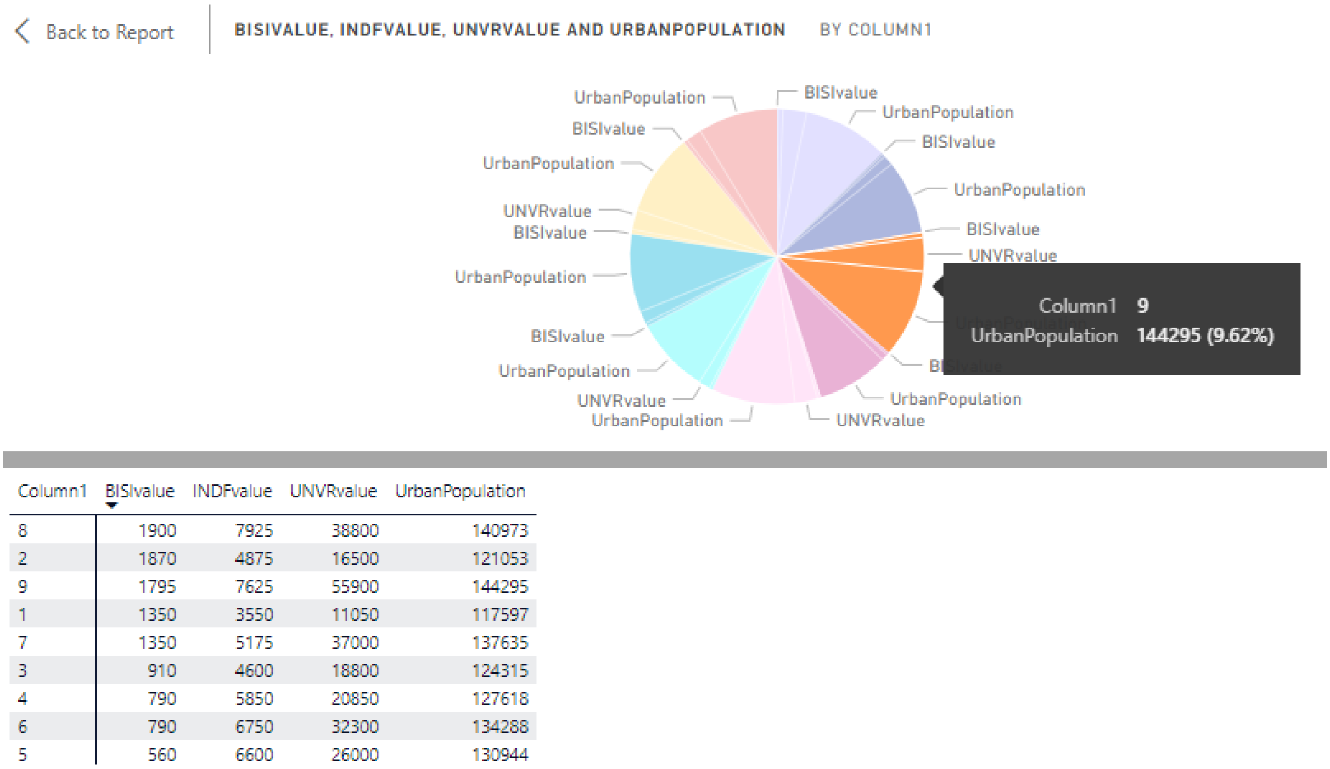


Figure 4: Results Of BI query2 from Power BI

7.3 BI Query 3: FMCG companies have upper hand in Indonesia?

This Business query was indirectly generated because of relationship between 1st two business query and to add more weightage to it data source from bloomberg.com and Statista completely clarified that its right business query. Stock data of three companies are considered and compared with urban poulation, which clearly show there is subtle amount in rise of stock value of fMCG companies. If urban population keeps on increasing at rate we are seeing in visualized form FMCG companies will have great amount of consumer base in Indonesia as gifted. Since demand of food will be easily matched by these industry in best and convenient way.

Visualized Business Queries are illustrated in Pie chart form



Column1	BISIvalue	INDFvalue	UNVRvalue	UrbanPopulation
8	1900	7925	38800	140973
2	1870	4875	16500	121053
9	1795	7625	55900	144295
1	1350	3550	11050	117597
7	1350	5175	37000	137635
3	910	4600	18800	124315
4	790	5850	20850	127618
6	790	6750	32300	134288
5	560	6600	26000	130944

Figure 5: Results Of BI query3 from Power BI

8 Conclusion and Future Work

This project main purpose is to design data warehose which will help in visualization of aggregate data in form of business intelligence, which in turn broaden scope of development and help in making efficient decision by government and FMCG industries. The visualization will also help in adding new data in incremental to check trial and error process easily. The data ware house complexity is reduced because of two dimension which defines whole data flow and new data generation. The data warehouse designed for this project can act a benchmark tool by relating critical factors to existing modern factor which impacts agriculture. Overall this process never end because data will keep on emerginng,adding,replacing to existing data warehouse process and simultaneously operation will happen in existing system. The complexxity data need to be handled in propered ways or else if dimension goes wrong it might be stressful to start from scratch.

Project can upgraded with lot of factors like region and rainfall month in dataset but data was not available which can relate to existing data warehouse raw data table, that's reason it was not considered. Since data analysis still predictive method tool by visualizing data one can only wait for automation of extraction of scrutinized data from website. Since two dimension is not enough to define any data warehouse project but these two dimension have clear impact on agriculture because analysis of agriculture is done by country government not by region and since agriculture takes long time to give desired result thats reason year is considered. The sole reason for choosing agriculture sector for datawarehouse project because it must need for today and the impact it can create in future on all industrial sector.

9 References

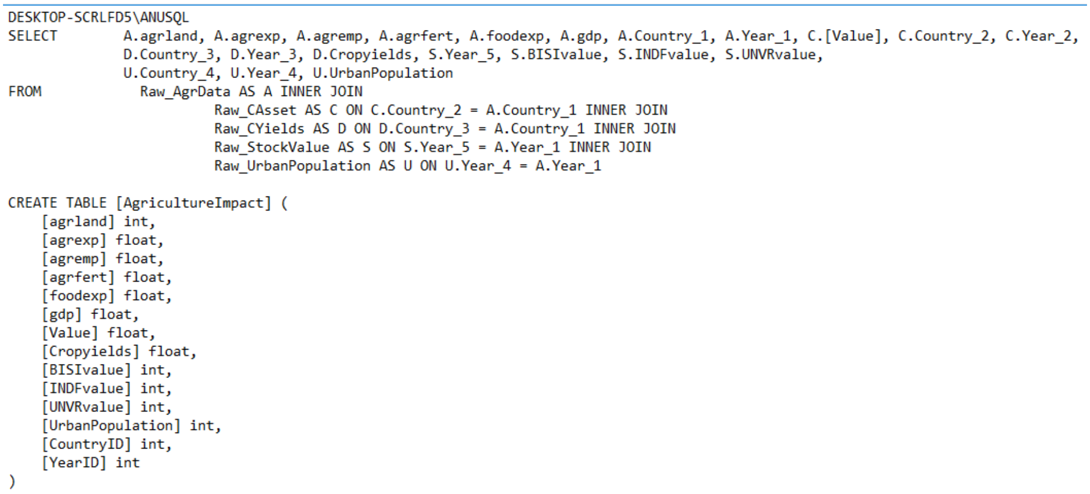
• Sustainable agriculture in Indonesia: Facts and challenges to keep growing in harmony with environment Syuaib, ”Sustainable agriculture in Indonesia: Facts and challenges to keep growing in harmony with environment”, Cigrjournal.org, 2018. [Online]. Available: http://www.cigrjournal.org/index.php/Ejounral/article/view/3747. [Accessed: 26- Nov-2018].

• Youtube

• Stackover flow

Appendix

9.1 sql query code

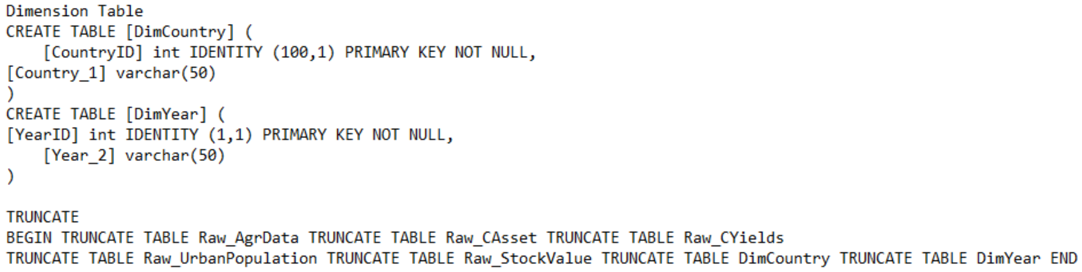


```
DESKTOP-SCRLFD5\ANUSQL
SELECT       A.agrland, A.agrexp, A.agremp, A.agrfert, A.foodexp, A.gdp, A.Country_1, A.Year_1, C.[Value], C.Country_2, C.Year_2,
             D.Country_3, D.Year_3, D.Cropyields, S.Year_5, S.BISIvalue, S.INDFvalue, S.UNVRvalue,
             U.Country_4, U.Year_4, U.UrbanPopulation
FROM           Raw_AgrData AS A INNER JOIN
                        Raw_CAsset AS C ON C.Country_2 = A.Country_1 INNER JOIN
                        Raw_CYields AS D ON D.Country_3 = A.Country_1 INNER JOIN
                        Raw_StockValue AS S ON S.Year_5 = A.Year_1 INNER JOIN
                        Raw_UrbanPopulation AS U ON U.Year_4 = A.Year_1

CREATE TABLE [AgricultureImpact] (
    [agrland] int,
    [agrexp] float,
    [agremp] float,
    [agrfert] float,
    [foodexp] float,
    [gdp] float,
    [Value] float,
    [Cropyields] float,
    [BISIvalue] int,
    [INDFvalue] int,
    [UNVRvalue] int,
    [UrbanPopulation] int,
    [CountryID] int,
    [YearID] int
)
```

Figure 6: SQL QUERY CODE

sql query code



```
Dimension Table
CREATE TABLE [DimCountry] (
    [CountryID] int IDENTITY (100,1) PRIMARY KEY NOT NULL,
[Country_1] varchar(50)
)
CREATE TABLE [DimYear] (
[YearID] int IDENTITY (1,1) PRIMARY KEY NOT NULL,
    [Year_2] varchar(50)
)

TRUNCATE
BEGIN TRUNCATE TABLE Raw_AgrData TRUNCATE TABLE Raw_CAsset TRUNCATE TABLE Raw_CYields
TRUNCATE TABLE Raw_UrbanPopulation TRUNCATE TABLE Raw_StockValue TRUNCATE TABLE DimCountry TRUNCATE TABLE DimYear END
```

Figure 7: SQL QUERY CODE

9.2 R Code

getwd()

Importing Raw Agriculture Data of Indonesia agr.data<-read.csv (”C:/Users/Dell/Documents/Indon Data.csv”,header=TRUE) agr.data

Clean Columns rows of Agriculture Data Of Indonesia agr.data1<-agr.data[-(7:13),c(2,5:13)] agr.data1 changing row to Year of agriculture Data names(agr.data1)[2:10] <-c(2009,2010,2011,2012,201 agr.data1 creating a new column of id of Agriculture dataset Agr$_id < -c(”agrland”, ”agrexp”, ”agremp$

Adding that column to dataset agr.data2=cbind(Agr$_id, agr.data1)agr.data2$

creating Matrix from dataframe and transposing it as.matrix(agr.data2) agr.data2.t=t(agr.data2) agr.data2.t

Transforming matrix to dataframe as.data.frame(agr.data2.t) agr.data3<-agr.data2.t agr.data3

Doing Modification in column and row by renaming it and deleting NA agrdata4<-agr.data3[-c(1:2),] agrdata4 colnames(agrdata4)[1]<-”agrland” colnames(agrdata4)[2]<-”agrexp” colnames(agrdata4)[3]<-”agremp” colnames(agrdata4)[4]<-”agrfert” colnames(agrdata4)[5]<-”foodexp” colnames(agrdata4)[6]<-”gdp” agrdata4 rownames(agrdata4)<-c(1:9) agrdata4 agr5<-agrdata4[-(10:49),] agr5

Creating previous column of country and year Year$_1 < -c(2009, 2010, 2011, 2012, 2013, 2014, 2015, 20$ $-c(”IDN”)Country_1$

Adding that column to dataset agr6=cbind(Country$_1, Year_1, agr5)agr6$

Importing Data from data frame to csv setwd(”C:/Users/Dell/Documents/Indonesia/DWBI”) write.csv(agr6,”Agrdata.csv”) write.csv(agr6) Importing Raw Crop Yields of Indonesia cyd <-read.csv(”C:/Users/Dell/Documents/Indonesia/DWBI/crop yields.csv”,header=TRUE) cyd Clean Columns of crop yields Of Indonesia cyd1<-cyd[,-c(2,5,8)] cyd1

Creating Dataframe for matrix df=cyd1[(1:9),5] df df1=cyd1[(10:18),5] df1 df2=cyd1[(19:27),5] df2 df3=cyd1[(28:36),5] df3 Creating Matrix of dataframe and adding year values

as.matrix(df) as.matrix(df1) as.matrix(df2) as.matrix(df3)

Cropyields=df+df1+df2+df3 Cropyields

again transforming matrix into dataframe as.data.frame(Cropyields)

Clean renaming Columns rows of Crop Yields Of Indonesia cy2<-cyd1[-(10:36),c(1:4)] cy2 names(cy2)[1]<-c(”Country$_3$”)$names(cy2)[4] < -c(”Year_3$”)

cy2[,2]<-c(”RWMR”) cy2 Adding calculated dataframe column to dataset cy3=cbind(cy2,Cropyields cy3

Importing Data from data frame to csv setwd(”C:/Users/Dell/Documents/Indonesia/DWBI”) write.csv(cy3,”Cyields.csv”) write.csv(cy3)

Importing Raw Cultivated Asset of Indonesia ca<-read.csv(”C:/Users/Dell/Documents/Indonesia/D Asset.csv”,header=TRUE) ca

Clean Columns of cultivated asset Of Indonesia ca1<-ca[,-c(2,5)] ca1 names(ca1)[1]<-c(”Country$_2$”)$names(ca1)[4] < -c(”Year_2”)ca1ImportingDatafromdataframetocsvsetwd(”C :$ $/Users/Dell/Documents/Indonesia/DWBI”)write.csv(ca1, ”Casset.csv”)write.csv(ca1)$

Importing Raw Urban Population of Indonesia up<-read.csv(”C:/Users/Dell/Documents/Indonesia/ Population.csv”,header=TRUE) up Clean Columns renaming of urban population Of Indonesia up1<-up[-c(1:6,16),] up1 names(up1)[1]<-c(”Year$_4$”)$names(up1)[2] < -c(”UrbanPopulation”)$ $-c(1 : 9)up1Country_4 < -c(”IDN”)Country_4Addingthatcolumntodatasetup2 = cbind(Country_4, up1$ $/Users/Dell/Documents/Indonesia/DWBI”)write.csv(up2, ”urbpop.csv”)write.csv(up2)$

READING BISI STOCK UNSTRUCTURED DATA FROM TEXT BISI1 <- read.delim(” /Indonesia/DWBI/BISI.txt”, header=FALSE, row.names=1, quote=””, stringsAsFactors=FALSE)

View(BISI1)

writing it to csv format setwd(”C:/Users/Dell/Documents/Indonesia/DWBI”) write.csv(BISI1,”BISI write.csv(BISI1)

reading csv file converting it to data frame B1 <-read.csv(”C:/Users/Dell/Documents/Indonesia/DWBI/BISI.CSV”,header = TRUE) B1 selecting stock value column B2<-B1[3] B2 BISIvalue<-B2[-(101),] BISIvalue giving it month name so data of december month can be taken into account month<-c(”MAY”,”JUN”,”JUL”,”AUG”,”SEPT”, ”OCT”,”NOV”,”DEC”,”JAN”,”FEB”,”MAR”,”APR”) month

modification in row and column data after identification of year B4=cbind(month,BISIvalue) B4 B5<-B4[-c(1:31,33:43,45:55,57:67, 69:79,81:91,93:103,105:115,117:127,129:140),] B5

adding year column to make data frame of BISI stock value Year$_5 < -c(2009, 2010, 2011, 2012, 2013, 2$

B6=cbind(Year$_5$, $B5)B6$

READING INDF STOCK UNSTRUCTURED DATA FROM TEXT INDF1 <- read.delim(” /Indonesia/DWBI/INDF.txt”, header=FALSE, row.names=1, quote=””, stringsAsFactors=FALSE) View(INDF1) writing it to csv format setwd(”C:/Users/Dell/Documents/Indonesia/DWBI”) write.csv(INDF1,”INDF.csv”) write.csv(INDF1)

reading csv file converting it to data frame I1<-read.csv(”C:/Users/Dell/Documents/Indonesia/DWI = TRUE) I1 modifying row columns of INDF excel I2<-I1[-(1:8),] I2 I3<-I2[2] I3 INDFvalue<-I3[-101 ,] INDFvalue

giving it month name so data of december month can be taken into account month1<-c(”DEC”,”JAN”,”FEB”,”MAR”,”APR”,”MAY”,”JUN”,”JUL”,”AUG”,”SEPT”,”OCT”,”NOV”) month1

making final data frame of INDF stock I5=cbind(month1,INDFvalue) I5 I6<-I5[-c(1:72,74:84,86:96,98:108,110:120,122:132,134:144,146:156,158:168,170:180),] I6

READING UNVR STOCK UNSTRUCTURED DATA FROM TEXT UNVR1<-read.delim(” /Indonesia/DWBI/UNVR.txt”, header=FALSE, row.names=1, quote=””, stringsAsFactors=FALSE) View(UNVR1) /newpage writing it to csv format setwd(”C:/Users/Dell/Doc write.csv(UNVR1,”UNVR.csv”) write.csv(UNVR1)

reading csv file converting it to data frame U1<-read.csv(”C:/Users/Dell/Documents/Indonesia/DW = TRUE) U1

modifying row columns of uNVR excel U2<-U1[-(1:8),] U2 U3<-U2[2] U3 UNVRvalue<-U3[-101 ,] UNVRvalue giving it month name so data of december month can be taken into account month2<-c(”JAN”,”FEB”,”MAR”,”APR”,”MAY”,”JUN”,”JUL”,”AUG”,”SEPT”,”OCT”,”NO month2 making final data frame of UNVR stock U5=cbind(month2,UNVRvalue) U5 U6<-U5[-c(1:71,73:83,85:95,97:107,109:119,121:131,133:143,145:155,157:167,169:179),] U6 combining all stock values SV=cbind(B6,I6,U6) SV STOCKVALUE<-SV[,c(1,3,5,7)] STOCKVALUE setwd(”C:/Users/Dell/Documents/Indonesia/DWBI”) write.csv(STOCKVALUE,”STO write.csv(STOCKVALUE)