**VIETNAM NATIONAL UNIVERSITY HCMC**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**FACULTY OF INFORMATION SYSTEMS**



**FINAL PROJECT**

**BI REPORT**

Lecturer: Nguyen Thi Kim Phung Msc.

Students: Duong Tran Son Long

Student ID: 175207006

June, 2022

**CONTENTS**

[1. **Introduction**. 5](#_Toc137025056)

[**a.** **Overview** 5](#_Toc137025057)

[*1.* *Background* 5](#_Toc137025058)

[*2.* *Dataset* 5](#_Toc137025059)

[*3.* *Datatypes* 7](#_Toc137025060)

[***b.*** ***Data warehouse when design*** 7](#_Toc137025061)

[**1.** **SQL Service Integrated Services Process** 9](#_Toc137025062)

[**a.** **Prepare the data** 9](#_Toc137025063)

[**b.** **Prepare the tools** 10](#_Toc137025064)

[1. Microsoft SQL server developer edition 2019. 10](#_Toc137025065)

[2. Microsoft Visual Studio 2019. (Package need: SSIS, SSAS) 10](#_Toc137025066)

[3. Power BI for Report 10](#_Toc137025067)

[**c.** **Start the process** 10](#_Toc137025068)

[***1.*** ***First, Create SSIS (SQL server + Visual code studio)*** 10](#_Toc137025069)

[***2.*** ***Second, do SSAS (visual code studio and SQL server)*** 14](#_Toc137025070)

[***a.*** ***Use Cubes query:*** 16](#_Toc137025071)

[***b.*** ***Use mdx query:*** 16](#_Toc137025072)

[***c.*** ***Use pivot excel query:*** 18](#_Toc137025073)

[***3.*** ***Third, Do SSRS With Power BI*** 19](#_Toc137025074)

[**4.** **Data Mining** 20](#_Toc137025075)

1. **Introduction**.
2. **Overview**
3. *Background*

Air pollution is a major problem in South Korea. On days with high pollution, citizens are advised not to go outdoors. This is especially true for those who are elderly or have pre-existing medical conditions. Pollution levels are higher at certain times of year and can change rapidly based on meteorological effects. Being able to accurately forecast the level of pollution would allow South Koreans to plan ahead and avoid exposing themselves to the harsh pollutants.

A picture containing text, sky, outdoor, water

Description automatically generated

1. *Dataset*

Pollution data

* date - date of measurement
* pm25 - fine particulate matter (PM2.5) (µg/m3)
* pm10 - fine particulate matter (PM10) (µg/m3)
* o3 - Ozone (O3) (µg/m3)
* no2 - Nitrogen Dioxide (NO2) (ppm)
* so2 - Sulfur Dioxide (SO2) (ppm)
* co - Carbon Monoxide (CO) (ppm)
* Lat - Latitude where measurement was taken
* Long - Longitude where measurement was taken
* City - City where measurement was taken
* District - District where measurement was taken
* Country - Country where measurement was taken*Data pre-processing*

*Link Dataset: https://www.kaggle.com/datasets/calebreigada/south-korean-pollution*

* 1. *Raw data*

*Graphical user interface, application, table, Excel

Description automatically generated*

1. *Datatypes*

|  |  |  |
| --- | --- | --- |
| *No* | *Attribute Name* | *Datatypes* |
| 1 | id | Int |
| 2 | Date | Date |
| 3 | Pm25 | Int |
| 4 | Pm10 | Int |
| 5 | O3 | Int |
| 6 | No2 | Int |
| 7 | So2 | Int |
| 8 | Co | Int |
| 9 | Lat | Float |
| 10 | Long | Float |
| 11 | City | Nvarchar (50) |
| 12 | District | Nvarchar (50) |
| 13 | Country | Nvarchar (50) |

1. ***Data warehouse when design***

***Graphical user interface

Description automatically generated***

* *Description Data warehouse:*

***At Table “Fact\_Pollution”:***

|  |  |  |  |
| --- | --- | --- | --- |
| *No* | *Attribute Name* | *Datatypes* | *Constraint Type* |
| 1 | id | float | PK |
| 2 | Pm25 | Int |  |
| 3 | Pm10 | Int |  |
| 4 | O3 | Int |  |
| 5 | No2 | Int |  |
| 6 | So2 | Int |  |
| 7 | Co | Int |  |
| 8 | Location\_id | Int | FK |
| 9 | Date\_id | Int | FK |

At Table “Gender”

|  |  |  |  |
| --- | --- | --- | --- |
| *No* | *Attribute Name* | *Datatypes* | *Constraint Type* |
| 1 | Location\_id | Int | PK |
| 2 | City | Nvarchar (50) |  |
| 3 | District | Nvarchar (50) |  |
| 4 | Country | Nvarchar (50) |  |

At Table “Product”

|  |  |  |  |
| --- | --- | --- | --- |
| *No* | *Attribute Name* | *Datatypes* | *Constraint Type* |
| 1 | Date\_id | Date | PK |
| 2 | Month | Int |  |
| 3 | Year | Int |  |

# **SQL Service Integrated Services Process**

## **Prepare the data**

As we know, we have raw data which contain the following information about shop sale.

Graphical user interface, application, table

Description automatically generated

## **Prepare the tools**

1. Microsoft SQL server developer edition 2019.

* Don’t use SQL server version 2019 will be error. Because it misses package “Pack version compatible with Microsoft Visual Studio 2019”

1. Microsoft Visual Studio 2019. (Package need: SSIS, SSAS)
2. Power BI for Report

## **Start the process**

1. ***First, Create SSIS (SQL server + Visual code studio)***

*Create a test database in SQL Server*

So firstly, we create database named **KoreanPollution\_DW** and **OriginalData** in SQL Server to store data from *south-korean-pollution-data.csv* file:

A screenshot of a computer

Description automatically generated with medium confidence

*Then we create new project SSIS from visual code studio 2019*

* Create SQL TASK + Dataflow for project

*A screenshot of a computer

Description automatically generated with medium confidence*

*At first Data flow task*

*A screenshot of a computer

Description automatically generated*

*At data flow task 2 we will bellow here*

*A screenshot of a computer

Description automatically generated*

*A screenshot of a computer

Description automatically generated with medium confidence*

*At data flow task 3 we do with Fact table*

*A screenshot of a computer

Description automatically generated with medium confidence*

*Don’t forget to add query for SQL Task:*

1. *Delete all data from all table.*
2. *Add Constraint foreign key and delete constraint foreign key.*

* Alter table Fact\_Pollution Add Constraint fk\_fact\_date Foreign key (Date\_id) References Dim\_Date (Date\_id);
* Alter table Fact\_Pollution Add Constraint fk\_fact\_location Foreign key (Location\_id) References Dim\_Location (Location\_id);

Then complete and run

When run complete and check database in SQL sever. Done !!

1. ***Second, do SSAS (visual code studio and SQL server)***

Create new project SSAS:

A screenshot of a computer

Description automatically generated

Then process cubes:

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence

Then test some cubes query:

### ***Use Cubes query:***

A screenshot of a computer

Description automatically generated

### ***Use mdx query:***

First login analysis database and write 15 query mdx:

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

### ***Use pivot excel query:***

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

1. ***Third, Do SSRS With Power BI***

First, connect Power BI with name Server and name DatabaseGraphical user interface, application

Description automatically generated

Wait, PowerBI will load data in sql server

A screenshot of a computer

Description automatically generated with medium confidence

1. **Data Mining**

*Purpose of Data Mining*

* After analyzing the data, the team exploits the attributes on the data warehouse to show the binding relationship and the relationship between the attributes.
* Use descriptive attributes of crime and crime level to make the final decision whether to be arrested or not.
* From those indicators, continue to learn, as well as do tests in the algorithms to find the input indexes that have the highest influence on the output, which is the Arrest capture attribute (True, False).
* Finally, after finding the appropriate input indexes, the team continues to use the decision tree algorithm - Decision Tree to make rules, compare with actual data to verify the accuracy of the decision tree. laws and make the best laws.

*There are 4 steps to do it:*

* Step 1: Create a new database, import data from excel file
* Step 2: Create new project, create data source, create Data source View
* Step 3: Create mining Structure (choose algorithm Decision - Decision Tree)

Select the predict . attribute

Select the input property

* Step 4: Execute the process, display the results.

Step-by-step process:

A screenshot of a computer

Description automatically generated

Graphical user interface, application

Description automatically generated