# Workshop 6:

# Data Warehouse Design Documentation

# For Singapore Dengue Insights

|  |  |
| --- | --- |
| Team Members: |  |
| Ho Kok Loon  Lee Tin Onn |  |
| Vincent Leung |  |
| Lei Jun |  |
| Choong Yue Lin  Peter Lee Wai Tong |  |

Contents

[1 Data Warehouse Objectives 3](#_Toc463817321)

[1.1 Introduction 3](#_Toc463817322)

[2 Data Warehouse Design 4](#_Toc463817323)

[2.1 Design Architecture 4](#_Toc463817324)

[2.2 Data Dictionary (Metadata) Design 6](#_Toc463817325)

[2.1 Data Entity Model 7](#_Toc463817326)

[2.1 Dimensional Modelling 8](#_Toc463817327)

# 1 Data Warehouse Objectives

## 1.1 Introduction

The objectives of the dengue data warehouse are to provide data insights into the possible factors that can affect the spread of dengue outbreak within Singapore in different time windows (daily, weekly and monthly). They include time sensitive data such as weather data, dengue clusters data and health data (More details can be found in later part of this document).

# Data Warehouse Design

## 2.1 Design Architecture



The above diagram depicts the overall hardware and software architecture of the data warehouse. In our data warehouse design, the ETL process basically takes place in the staging area where the externally sourced data are located.

The dengue data, weather data and health data are all extracted externally from websites such as data.gov, weather.gov using web scraping scripts and form into their own respective data groupings. They are subsequently subjected to the ETL process at the Data Acquisition Later for further extraction and transformation before loading into the data warehouse proper.

Inside the data warehouse, the ETL process will generate the enterprise data warehouse (depicted as raw data in the above diagram) and subsequently data marts (depicted as summary data) get created from the larger data warehouse. It is in this data mart (summary data) where data gets broken into smaller units to cater for different use cases. In our case, we are generating data to allow users to have insights on what could be possible factors that can affect dengue outbreak both from individual factors such as weather data (PSI, Temperature, UV Index etc.) broken into different time frame.

Below are the processes involved in the ETL process: It is in this data mart (summary data) where data gets broken into smaller units for a flat data model to cater for different business user group needs.

1. Perform data extraction function to extract data from external sources (in our case there are mainly the data.gov and weather.gov websites) regularly. In this data warehouse design, the approach is to simply generate Comma Separated Value (csv) files for each of the data groupings listed down.
2. In the process of data extraction, pre-processing also takes place to process hourly data information into daily data as well as to discard data which has incomplete data for the key fields. The hourly data gets processed into derived data such as max, min and average data depending on the suitability of the data for the business user groups.
3. As the database volume is not big for this assignment, the output of the database is in the form of ETL.

## 2.2 Data Dictionary (Metadata) Design

The excel file attached in this document documents the data dictionary design. Click this [link](Data%20Dictionary.xlsx) for the details.

(Note if the link could not be opened, please open the <Data Dictionary.xls> file to view the data dictionary design).

Below is a summary description of the data dictionary which forms part of the metadata management. As the data warehouse is not complex, we did not go ahead to develop the full metadata that includes other aspect of the data warehouse such as names of stored procedure associated with the each of the databases etc.

## 2.1 Data Entity Model

Below shows the data entity model of the Data Warehouse depicting the various relationships between each of the entities. <WeatherStation> entity which stores the location information of the has a 1 to many relationship with the weather. <Region> entity has a 1 to many relationship with <WeatherStation> entity. <Region> also has a 1 to many relationship with <DengueDisease>, <UVIndex>, <PSI> entities.



NOTE:

1. The data is daily based, so ETL should do some calculation accordingly.
2. For some data like UV index, need to calculate the max/mean/min Index based on current hourly raw data during ETL;
3. Make very table a file since all the data will be stored in csv files;
4. The region table should include all the regions in Singapore, and also one record is the overall Singapore without specified region, for those table who doesn’t have region information, which means the general Singapore area.

## 2.1 Dimensional Modelling

The below shows the Dimension table design (after all the prior normalization process). Items starting with dim is the Dimensional table and items starting with fact is the Fact table. The relationships between the tables via primary and foreign key are also shown here.

