LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY
School of Engineering Science
Software Engineering

Grace Heinonen

DATA WAREHOUSE FOR CLIMATE-BASED INFECTIOUS DISEASE SURVEILLANCE SYSTEM IN FINLAND

Examiners: Professor Ajantha Dahanayake
Jiri Musto, M.Sc. (Tech.)

ABSTRACT

Lappeenranta-Lahti University of Technology
School of Engineering Science
Software Engineering
Grace Heinonen

Data Warehouse for Climate-Based Infectious Disease Surveillance System in Finland
Master's Thesis 2019

68 pages, 17 figures, 15 tables, 1 appendix
Examiners: Professor Ajantha Dahanayake
Jiri Musto, M.Sc. (Tech.)

Keywords: data warehouse, disease surveillance, climate and infectious diseases

Infectious diseases often occurred as an epidemic, placing human mortality and morbidity at great risk. It is scientifically proved that the life cycle and transmission of many infectious disease pathogens are inextricably intertwined with climate. The surveillance system possesses a significant role in controlling the diseases, pathogens and its clinical outcome. Considering the sensitivity of climate towards infectious disease, it is reasonable to set climate parameters such as weather and air quality as predictive indicators of the disease surveillance system. Supported by rapid development in big data, this breakthrough is feasible from the technical point of view. Explosive growth and accessibility of digital healthcare data, however, possess another issue. They are not only big in volume but also complex in variety and high at velocity. The incapability to deal with these traits concludes that the traditional database system is not the right equivalent for the big data stream. A powerful solution, therefore, is critically needed. This thesis work will explicate the design of a data warehouse as the proposed solution for the above problem.

ACKNOWLEDGEMENTS

This thesis work would never happen without Professor Ajantha Dahanayake. To her, I express my most sincere gratitude for the indispensable guidance on this thesis work and the whole years of my master's study at LUT.

Thanks to my whole family in Indonesia and Finland for their continuous prayer and encouragement allowing me to come this far. And of course, Joona, the best companion I could wish for this peculiar adventure.

Finally, thank God that I am writing this very last part with such a big smile.

Lappeenranta, 1 December 2019

Grace Heinonen

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

BI	Business Intelligence
CO	Carbon monoxide
EDW	Enterprise Data Warehouse
ETL	Extract Transform Load
HOLAP	Hybrid Online Analytical Processing
ICD	International Classification of Diseases
IDC	International Data Corporation
IT	Information Technology
MOLAP	Multidimensional Online Analytical Processing
NO_2	Nitrogen dioxide
O_3	Ozone
OLAP	Online Analytical Processing
PM	Particulate matter
ROLAP	Relational Online Analytical Processing
SLR	Systematic Literature Review
SO_2	Sulphur dioxide
WHO	World Health Organization

1 INTRODUCTION

This chapter provides description of background, objectives and delimination, research methodology and structure of this thesis work.

1.1 Background

Infectious diseases, or communicable diseases, are health disorder engendered by pathogenic agents, transmitted by person or animal. World Health Organization (WHO) referred to this problem as a great burden for many societies. As it often occurred as epidemics, the infectious disease threatens human mortality and morbidity on the national or even higher scale. Over twenty-five centuries ago, Hippocrates and his predecessors have discussed the impact of climate change on infectious disease. They stated that it is caused by a disturbance in the part of the constituent body which triggered by atmospheric and climatic conditions. Nowadays, many scientific works have demonstrated the truth of this statement; there are many well-documented data and findings indicates the correlation of climate parameters such as weather and air quality with infectious diseases occurrences (Hippocrates and Jones, 1923; WHO, 2001, 2005; P. Polgreen and E. Polgreen, 2017).

The rapid improvement in big data has increased the interest of scientific works in healthcare and Information Technology (IT) areas. Commonly known as computational health informatics, this topic has become an interesting object of research among IT researchers over the past few years. It is a multidisciplinary field that covers several subspecialties including clinical research informatics. The primary focus is data warehouses development healthcare research. Disease surveillance is considered as one of the most interesting subjects in this domain. It is used to monitor the emerging and re-emerging of infectious diseases, for example, respiratory infections, gastrointestinal infections and antimicrobial resistance in some specific areas. Disease surveillance provides a mechanism for government and healthcare organizations to report, detect and prevent the dissemination of infectious diseases on time to minimize the detrimental effects on the human population. It is an important pillar to control the disease, pathogen, and its

outcomes. Finally, the utilization of disease surveillance also worthwhile as it produces all the necessary information about the potential factors which trigger specific circumstances (Fang et al., 2016; Bansal et al., 2016).

Consider the linkages between climatic conditions and infectious disease occurrences, it is reasonable to use the climate parameters as the indicator to predict the growth of the epidemics for different purposes. This is including the development of a surveillance system to control the disease. Furthermore, this analysis process becomes more feasible from the technical perspectives. This progress was mainly affected by the improvement in the availability of climate and healthcare data as well as the use of geographical information systems and remote sensing (WHO, 2005).

In contrast with the conventional approach, healthcare data nowadays are electronically generated with automatic devices such as mobile phones, wearable devices, radio-frequency sensors, and satellites. This invention leads to the healthcare big data stream. During 2013, the estimation of global healthcare data was almost 153 exabytes; it is even predicted to reach 2314 exabytes in 2020 which equates to a 48% annual increment. This explosive growth and widespread accessibility of healthcare data, unfortunately, raises new issues (Fang et al., 2016; Stanford Medicine, 2017). Table 1 shows the challenges of integrating big data with disease surveillance.

Table 1. Challenges of integrating big data with disease surveillance (Hay et al., 2013)

Big Data Characteristic	**Occurrence Point**	**Pseudo-Absence Point**	**Environmental Covariates**	**Risk Prediction**
Volume (scale)	High	Low	High	High
Velocity (frequency)	High	Medium	High	High
Variety (diversity)	Medium	Low	Low	Low

Healthcare data that automatically generated through big data streams are not only huge in amount. As seen in **Table 1**, they also complex in structure and high at speed. Traditional database systems with poor time lags and unavailability of spatial resolution are not compatible with big data. It is incapable to extract information from the massive and disorganized data stored in different repositories and transmitted at high speed (Khan and Hoque, 2015; Bansal, 2016). A powerful solution is in need. This research is heavily focused on the design of a data warehouse as the proposed solution for this problem.

1.2 Objective and Delimitations

The objective of this thesis work is to design a data warehouse that integrating climate data, specifically weather, air quality, and marine observation data with infectious disease register data. It is intended to support the infectious disease surveillance system based on climate conditions in Finland. These four main datasets were taken from the Finnish Meteorological Institute and Finnish Institute for Health and Welfare. This project will focus on the data warehousing part without going any further to the data mining part.

Furthermore, the delimitations of this research are formulated into two main questions:

a. What are the roles and contributions of the data warehouse for healthcare informatics and public healthcare in general?
b. How to integrate infectious disease register data with weather observation data, air quality observation data and marine observation data into one single repository without overriding its consistency?

1.3 Research Methodology

Systematic Literature Review (SLR) is used to identify and evaluate the available material related to the utilization of data warehouse for climate-based disease surveillance systems. The literature used for this thesis work was strictly selected according to its relevance with the discussed topic to confirm that information is collected objectively and qualified to be used as a proper base to form insights into the research areas.

Table 2 provides the number of scientific literature found on different academic databases according to keywords of this research: data warehouse, disease surveillance (system), climate and infectious diseases. As shown, a considerable amount of literature sources were devoted to these specific areas. However, there are still spaces for research to relate these three particular keywords into one single domain.

Table 2. Number of scientific literatures for different keywords found on academic databases

Scientific Database	**Keywords**		
	Data Warehouse	**Disease Surveillance**	**Climate and Infectious Disease**
ACM Digital Library	182,331	351,015	4,547
IEEE Explore	4,282	547	21
SAGE Journals	9,918	33,341	9,409
ScienceDirect	43,449	173,242	25,412
Springer Link	60,959	157,031	23,202

The design of the data warehouse in this thesis work follows the framework proposed by Vaisman and Zimányi (2014). It consists of four phases. The first phase is the requirement specification to define the essential elements of the system and how it should be organized. The second phase is conceptual design to build the conceptual schema for the data warehouse which represent a set of the requirement clearly and concisely. The third phase is the logical design to convert the conceptual schema into the logical schema and specify the staging and Extract, Transform and Load (ETL) process. The fifth phase is physical design to convert the logical schema into the physical data warehouse structure.

1.4 Structure of the Thesis

Following this introduction part, the rest of this thesis work is divided into five chapters as

follows. The second chapter is the literature review, presents the previous work related to this topic as the foundation of this thesis work. The third chapter is the design methodology, clarifies the theoretical background of data warehouse design methodology which used in this research. The fourth chapter is data warehouse design, elucidate each phase of data warehouse design for this work following the framework as explicated on the third chapter. The fifth chapter is the discussion and description of further work needed relating to this topic. The sixth chapter is the conclusion of the research.

2 LITERATURE REVIEW

This chapter provides the literature review related to the data warehouse, disease surveillance system and its sub-field which relevant to the topic of this research.

2.1 Data Warehouse

The term of the data warehouse was firstly introduced in the 1980s. It was developed as an alternative to store and organize data in a consolidated and integrated manner for statistical analysis and Business Intelligence (BI) purposes (Salinas and Lemus, 2017). It is a phenomenon that arises because of the enormous amount of digital data stored during recent years and the urgency to use these data to support the organization (Golfarelli and Rizzi, 2009). The motivation of developing a data warehouse is formed by three primary needs: the need of business for global and independent view of information, the need of information system to organize big data more effectively and the need of reducing reporting load as well as operational database servers (Foster and Godbole, 2016).

This literature review part describes data warehouse mainly from the perspective of two important figures and pioneers of this area: Bill Inmon, regarded by many as the father of data warehouse and Ralph Kimball, one of the initial architects of the data warehouse. Additionally, reviews from scholars and professionals were also considered.

2.1.1 Characteristics

Inmon (2002) define data warehouse as “a subject-oriented, integrated, non-volatile and time-variant collection of data in support of management’s decision” (p. 31). The data warehouse is subject-oriented because the operating systems are based on enterprise-specific applications. The application for an insurance company possibly consists of auto, health, life, and casualty with the major subject area of customer, policy, premium, and claim. The data warehouse is integrated; data contained therein are generated from

different and divergent sources to be converted, re-formatted, re-sequenced, summarized, etc. The data warehouse is non-volatile, it allows alteration, but the update is loaded in a snapshot static format to ensure that the history of data is stored thoroughly. The data warehouse is time-variant; the normal collective time for data horizon in a data warehouse is between 5 to 10 years. Fundamentally, data is never been omitted from the data warehouse repository; it is regularly updated from source systems and keeps on growing. (Golfarelli and Rizzi, 2009).

Furthermore, Inmon also introduced twelve standards for managing the data warehouse. First, the data sources and data warehouse must be separated. Second, the data warehouse is an integration of different source systems. Third, the data warehouse typically contained historical data collected during an extended period. Forth, the nature of data must represent a snapshot of the operational data source at a specific time. Fifth, the contained data are subject-oriented. Sixth, the contained data are predominantly read-only databases which updated periodically from connected operational databases. Seventh, the life cycle of the data warehouse is data-driven. Eighth, there is a possibly different level of detail, for example, current detail, old detail, lightly summarized detail, and highly summarized detail. Ninth, data set are characterized by a read-only transaction on big data set. Tenth, the data warehouse must pose a system to keep track of all data sources, transformation, and storage. Eleventh, metadata forms should be able to provide functions of the definition of data elements, identification of data source, transformation, integration, storage, usage, relationship, and history of each element. Twelfth, the most optimal resource usage of data is reached by applying the different form of chargeback mechanism (Foster and Godbole, 2018).

On the other hand, Kimball and Ross (2010) describe the Enterprise Data Warehouse (EDW) as “the union of a set of separate business process subject areas implemented over a period of time, possibly by different design teams, and possibly on different hardware and software platforms” (p. 210). There are six requirements for a data warehouse. First, the content of the data warehouse must be comprehensible and valuable. In addition to that, the used tools must also be simple, easy and able to return the query result with minimum processing times. Second, it must be able to represent the information consistently. In other

words, the provided data must be reliable. Third, the data warehouse must be adaptive and resilient to any alteration. It must be able to handle the inevitable changes without invalidating the existing data or application. Forth, the data warehouse must be a secure bastion that protects the information assets. It must be able to control access to organizational confidential information. Fifth, the data warehouse must be able to serve as the ground of decision making as it is fundamentally targeted to support decision making. Sixth, the business community must implement the data warehouse to be considered successful. Unlike the operational system rewrite which required to be used, data warehouse usages sometimes work as an option. Therefore, it is important to make sure that the business community actively uses the data warehouse after the training period (Kimball and Ross, 2013).

2.1.2 Components and Architectures

These five traits are substantial for a data warehouse. Firstly, separation; the analytical and transactional system must be separated as far as possible. Secondly, scalability; hardware and software architecture must be easily upgradeable according to the data volume and user requirements which are progressively increasing. Thirdly, extensibility; architecture must be capable of executing new applications and technology without altering the current system. Fourthly, security; access monitor is important considering all the strategic data stored in the data warehouse. Fifthly, administrability; management of data warehouses should be convenient (Golfarelli and Rizzi, 2009).

Various architecture models for the data warehouse are commonly recognized in much-related literature. However, this thesis work will only highlight specific architectures according to two distinctive and prominent approaches in this area: the top-down approach and the bottom-up approach. These two architectural approaches practically consisted of similar components and functions. The key difference is in the method used for modeling, loading and storing the data into the data warehouse. This method will affect the initial preliminaries of the data warehouse design and the capacity of resettling any design transformations in the future (Rangarajan, 2016).

The top-down approach was proposed by Inmon. With this approach, the data model is specified in the initial phase and serves as a foundation to identify the key subject and entities of the business such as product, supplier, and customer. The data warehouse is a centralized repository where the atomic data are stored at the lowest level and built before data marts. The examples of data warehouse architecture with the top-down approach are centralized and hub-and-spoke architecture (George, 2012; Rangarajan, 2016). Illustration of data warehouse architecture with the top-down approach is shown in **Fig. 1**.

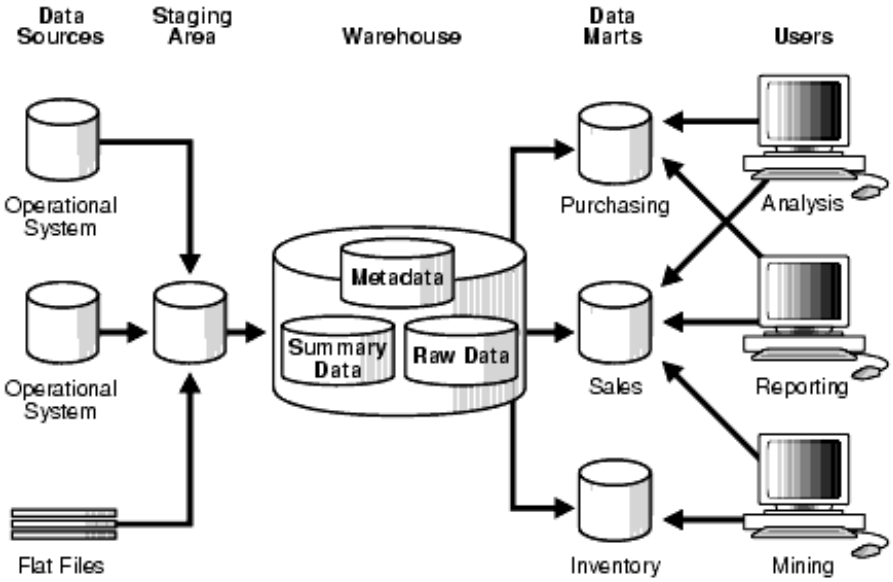

Fig. 1. Inmon data warehouse architecture (Lane, 2005)

The above architecture consists of the following components. The Data sources are generally divergent. It can be fetched from the business information system such as the operational database and flat files and also reside from outside of the company. Data staging is an intermediate component. It lies between the data sources and data warehouse, where data are integrated and transformed to be loaded into the data warehouse. This whole process commonly known as the Extract, Transform and Load (ETL) process. The data warehouse is the centralized repository to store enterprise data in a multidimensional form. It has a metadata repository to describe the structure of the data warehouse at different levels, security and monitoring information, data sources and its schemas, and the ETL process. The data warehouse serves as the source to create several data marts. These are the specialized repository, storing specific parts of the information from the data

warehouse which relevant for particular necessity inside the business. Data marts can be accessed by the user with different types of tools for analysis, reporting and mining purposes. (Rizzi, 2008; Golfarelli and Rizzi, 2009; Vaisman and Zimányi, 2014).

The bottom-up approach was introduced by Kimball. Contrary to the top-down approach, different data marts are created first based on the characteristic of related business processes or areas. These then will be merged to create a global data warehouse (George, 2012). Independent data mart architecture and bus architecture are examples of data warehouse architecture constructed with the bottom-up approach. The illustration of data warehouse architecture with the bottom-up approach is shown in **Fig. 2**.

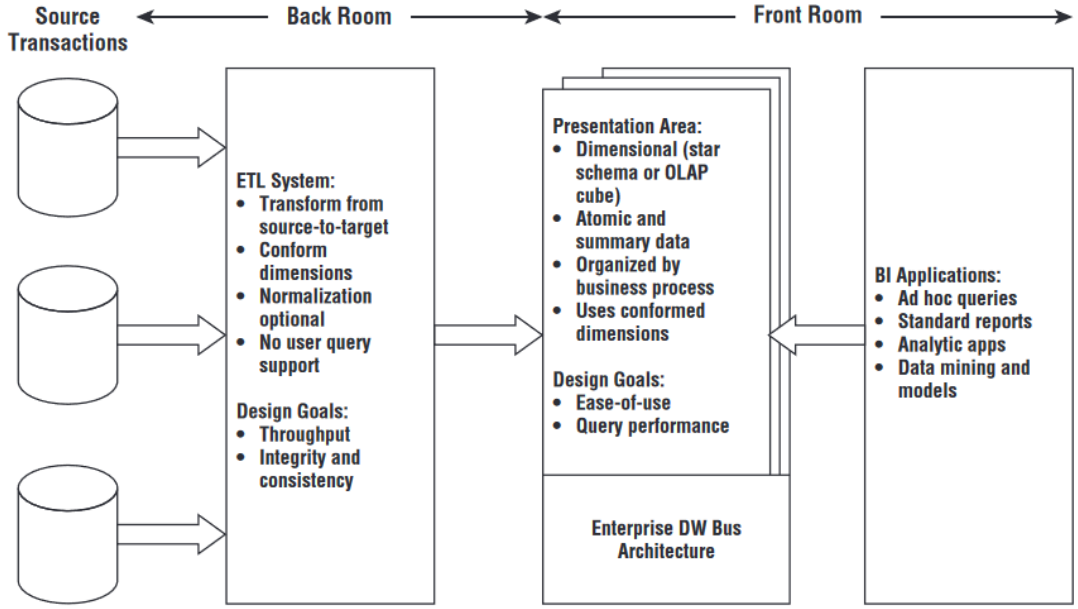

Fig. 2. Kimball data warehouse architecture (Kimball and Ross, 2013)

There are four separate and distinct components serves specific functions. Source systems serve as the recorder to capture the business transaction with the main priorities to processing performance and availability. ETL system separates the source systems from the presentation area; it functioned to adds value to the extracted data by transforming it with numerous processes such as cleaning, integrating data from different sources, deduplicating data, and defining the primary key. This whole process must be completed before retrieving the data into the presentation area. Data presentation is responsible for organizing and presenting the stored data requested by the user for various applications

such as reporting and analyzing. The term of BI application loosely refers to different capacities of users to utilize the presentation area for different analysis processes to make a decision; it can be easy as an ad hoc query tool or difficult as a complex data mining or modeling application (Kimball and Ross, 2002; 2013).

To decide which approach to be used, several considerations must be taken into account. This includes the qualifications of the team, the requirement specification of the intended data warehouse, and the financial support (Vaisman and Zimányi, 2014). Comparisons of these two approaches are listed in **Table 3**.

Table 3. Comparisons of the top-down and bottom-up approach (George, 2012)

Factor	**Top-Down Approach**	**Bottom-Up Approach**
Duration	Time consuming	Take lesser time
Maintenance	Easy, low redundancy and flexible	Difficult, high redundancy and subject to revision
Cost	High preliminary cost with lower sub-sequent processes cost	Low initial cost with similar subsequent processes cost
Initial process	Longer initial process	Shorter initial process
Skill	Specialist team	Generalist team
Scale	Enterprise-wide	Individual business area

The development of a data warehouse with the top-down approach might be challenging considering its duration, cost, size, and complexity. However, it produces a fully functionates system that works as a single source of truth for all data marts in the whole enterprise. As the data redundancy is extremely low, anomalies can be avoided, the ETL process is more convenient and less susceptible to failure. On the contrary, the bottom-up approach generally faster in delivering the data mart at a lower cost. Nevertheless, the concept of one single source of truth vanishes because there is no more full integration among each data. Data is redundant and potentially caused to anomalies. The development of data marts requires global frameworks to ensure future integration. Lack of proper

structure will result in difficult processes and high costs in the long term (Vaisman and Zimányi, 2014; Rangarajan, 2016).

2.2 Diseases Surveillance

WHO (2006) define surveillance as "the ongoing systematic collection, analysis, and interpretation of outcome-specific data for use in planning, implementing and evaluating public health policies and practices" (p. 1). Furthermore, disease surveillance must be able to serve these two primary functionalities. Firstly, the early warning of threats that potentially harmed public health. Secondly, monitoring function, be it diseases-specific as well as multi-diseases in nature. Disease surveillance provides monitoring functions over time and depending on nature. It may be an appropriate instrument for detecting unusual patterns among the data. Disease surveillance possesses detecting features rather than predicting the epidemic onset. However, it can be used to support early warning system if it is collected and analyzed routinely (WHO, 2005). An example of the surveillance system and its indicators were listed in **Table 4** below.

Table 4. Example of surveillance system (WHO, 2015; Groseclose and Buckeridge, 2017)

Surveillance System	**Purpose**	**Area of Investigation**	**Data Source**
Italian behavioral risk factor surveillance system (PASSI)	Observe health behavior and the risk factors to support health promotion and disease prevention on the local level	Alcohol consumption, diet, and nutritional status, physical activity	Telephone-based interview every month
Swedish Västerbotten Intervention Programme	Reducing morbidity and mortality caused by cardiovascular disease and diabetes	Various disease outcomes, demographic and socioeconomic conditions assessment	Self-reported health, laboratory measurement

WHO European Childhood Obsesity Surveillance Initiative (COSI)	Monitor the prevalence of biological risk factor for non-infectious diseases like overweight and obesity in children	Measure trends of overweight and obesity in primary-school-age children on a routine basis	Children's weigh measurement
Moldovan National STEPwise Approach to Surveillance (STEPS)	Evaluate the prevalence of main infectious disease risk factor for more efficient prevention plan, control policies and activities	Sociodemographic, behavior, physical and biochemical	Sociodemographic and behavioral information, physical and biochemical measurement
Gonococcal Isolate Surveillance Project	Supports the early detection and response towards outbreaks and new emergency health condition	Monitoring the trends of *Neisseria gonorrhoeae* among men with gonococcal urethritis case who visiting one of the 27 special clinics in the US	Clinic visit data
Active Bacterial Core Surveillance System	Evaluation of public policy for developing vaccine guide and immunization policy	Active network surveillance in 10 geographically and racially diverse jurisdiction up to 12% of US population	Active surveillance data

Several disease surveillances are used for the surveillance stream to examine the impact of climatic changes towards the geographic distribution of disease. A Surveillance system consisted of two main components. First is the indicator-based surveillance for monitoring the frequency, origin, and distribution of reportable disease. It uses data structured based

on the case definition, generated by outpatient consultation and inpatient admission cases, for example, mortality data, morbidity reports, laboratory data, vaccine, and drug utilization. Second is the event-based surveillance for recognizing events by actively scanning the internet media and sources of big data, making ad-hoc contact with health providers and the community to detect potential risk. It uses data which not necessarily follow the specific case definition. However, the unstructured data will be analyzed first to determine the presence of risk factors, for example, sickness absence data from school or workplace and weather forecast (WHO, 2005; Greenwell and Salentine, 2018). **Fig. 3** below illustrate an example of a framework for specifically developing a climate-sensitive disease surveillance system.

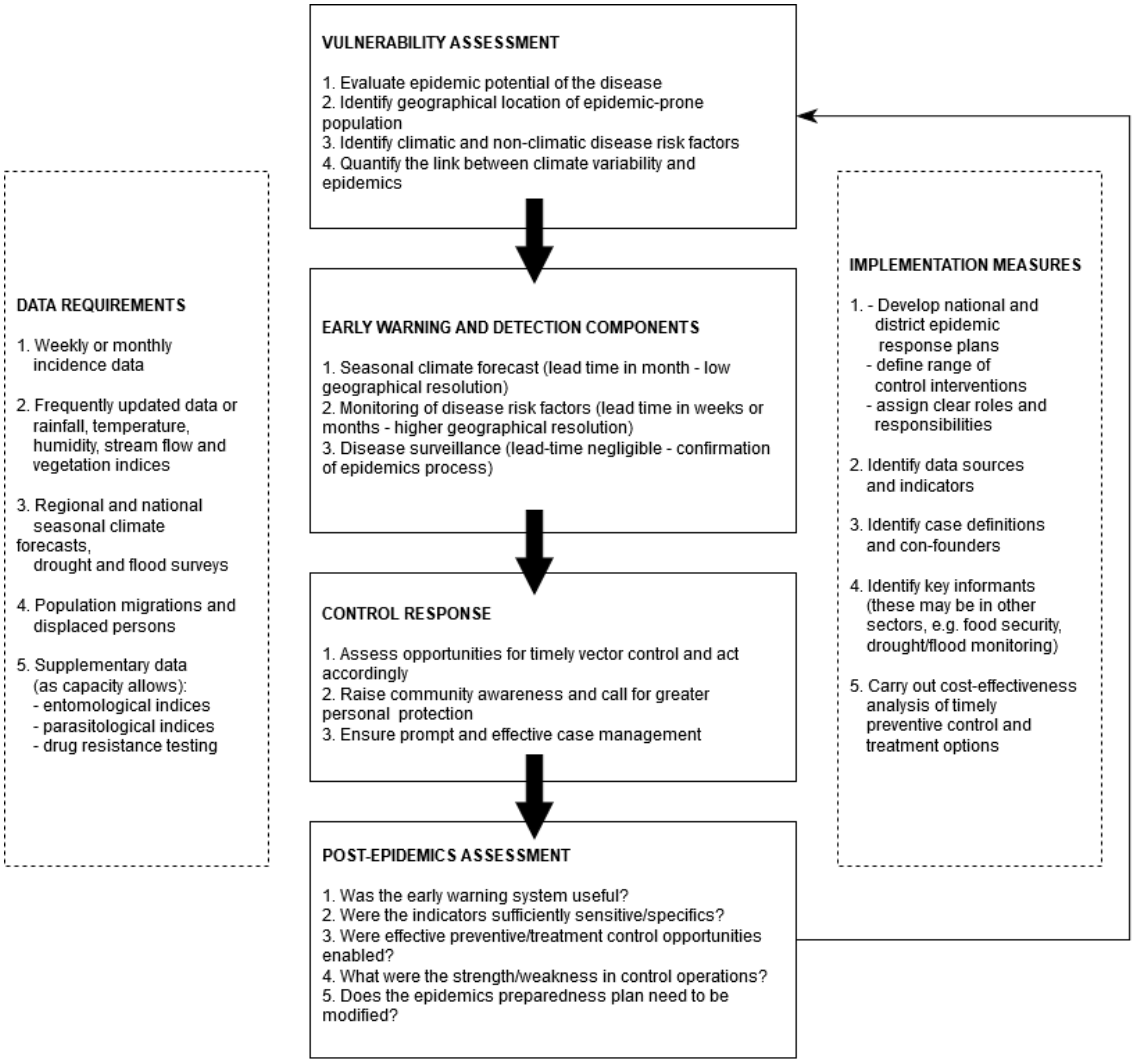

Fig. 3. Framework to develop surveillance system for climate-sensitive diseases (WHO, 2005)

This framework composed of four preliminary phases: evaluation of potential factors for epidemic transmission which are varying and markedly by the season, identification of the epidemics geographical areas as well as climatic and non-climatic variables. The next and last phase is the measurement of the linkage between climate variance and disease occurrence by constructing the predictive model (WHO, 2005).

A surveillance system is a representation of the population, flexibility, economic condition, and social resilient reported and validated on time. It required multiple data streams that captured mild and severe clinical outcomes as well as laboratory-based information. Surveillance systems not only used for detecting the fluctuation of health threats. It can also be used to detect the increase of outbreak frequency, change of seasonal incidences, risk of geographical distribution, and new pathogens or disease vectors in some specific areas. Surveillance systems provide important information to find the relationship between different parameters like climatic, environmental, socio-economic, and demographic conditions to discover the caused of these specific circumstances (Nichol et al., 2014; Simonsen et al., 2016). **Table 5** listed the history of the disease surveillance system.

Table 5. History of disease surveillance and big data (Simonsen et al., 2016)

Year	**Diseases**	**Initiator/ Disease Surveillance**	**Institution (s)**
1662	Plague	John Graunt	Bills of Mortality, UK
1817	Smallpox	J.C. More	
1847	Influenza	William Farr	General Registrar Office, UK
1854	Cholera	John Snow	General Registrar Office, UK
~1990	All	International Classification of Diseases (ICD)	WHO
1918	Influenza	121 Cities Mortality Reporting System	Centers for Disease Control and Prevention (CDC)
1976	Influenza	Weekly viral surveillance	CDC

1984	Influenza, viral hepatitis, acute urethritis, measles, mumps	Introduction of the computerized disease surveillance network	French Sentinelles Network
~2000	All	Introduction of medical claims forms	Private companies and government
2008	Influenza	Launching of Google Flu trends	Google
2015	Influenza	Birth of hybrid systems	Public/private partnership

The development of disease surveillance systems began around the 19th century in Europe and North America. At that time, the incidents of diseases and deaths were reported by doctors and physicians every week. Based on this report, some important decisions related to healthcare policy were made, for instance, introducing the smallpox vaccination program and further evaluation for intervention purposes. Nowadays, these classical surveillance system has experienced such a tremendous improvement. Relying on the big data stream, it covers death certificates, patient records, and medical claims which re-organized according to the standard of ICD to compare the pattern of the syndromic disease during a specific time in a specific area. In line with that, utilization of digital data generated from social media and crowdsourcing for surveillance systems has been introduced and is in use (Simonsen et al., 2016).

2.2.1 Relating Climate with Infectious Diseases

Far before the notion of infectious pathogens were discovered in the nineteenth century, human has already known that weather changes and climatic condition affect the epidemics occurrences. The Roman aristocracy used to take shelter in their hill resort during summer to protect themselves from malaria. People in South Asian even purposely consume strongly curried food to induce diarrhoeal diseases in summer. Back in 1878, a yellow fever outbreak occurred during summer in the southern United States. This incidence was considered one of the most terrible outbreaks and it happened during one of the strongest

El Niño in history, causing an enormous disaster with an estimated death toll of around 20000 people. Nowadays, it is commonly accepted in developed countries that recurrent influenza epidemics happened during mid-winter (Patz et al., 2003).

Kristie et al. (2017) stated that the connection of climate variability to infectious disease and human health often complex and indirect. It multiplies the stress and adds even more pressure on the vulnerable system, population, and area. Climate change affects the infectious disease through changes in the frequency and intensity of weather, air and water quality. Temperature, for example, often associated with the occurrence of several food-and-water-borne diseases causing children mortality. **Fig. 4** depicted the connection of climate variability to infectious diseases and human health.

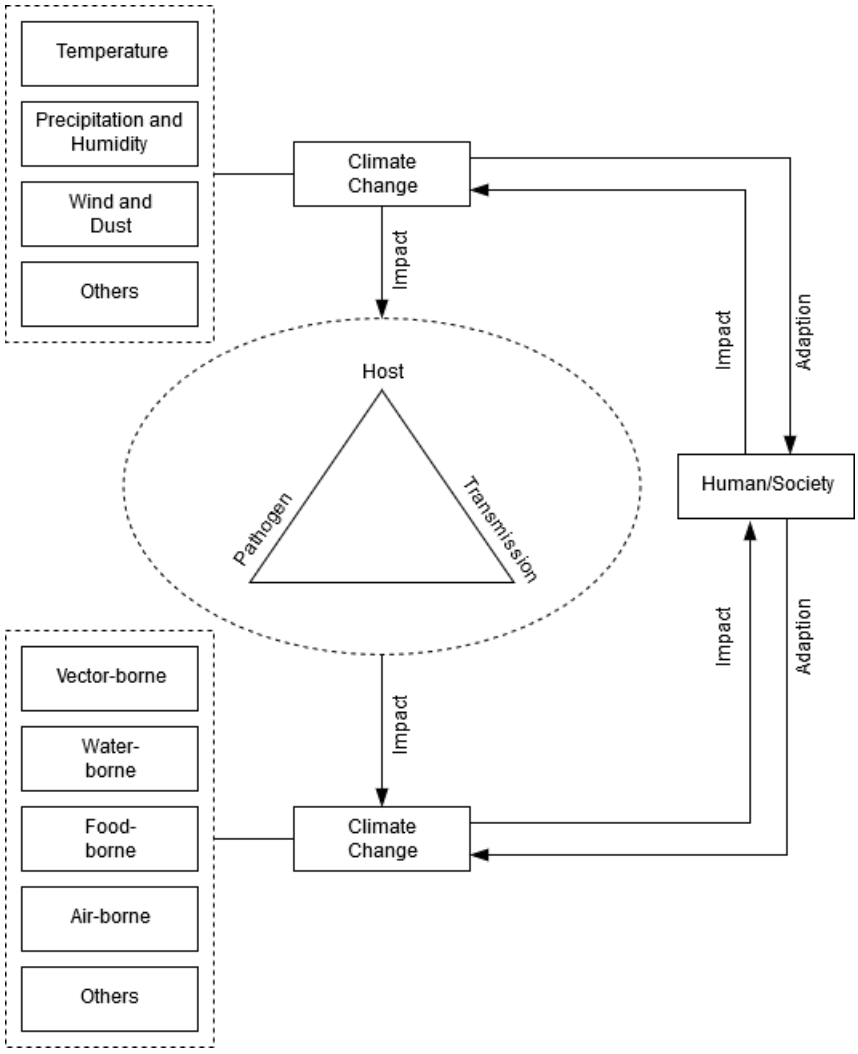

Fig. 4. Connection of climate variability to infectious disease and human health (Wu et al., 2015)

Many studies have provided scientific evidence that the distribution of infectious diseases is inherently linked to climate change. These diseases often occur as epidemics, risking the mortality and morbidity of the human population. Although this climate change is a global phenomenon, it engenders significantly different effects on human health. These effects vary from the obvious risk of extreme temperature and malignant storms to the unclear and unpredicted correlation. Climate change also has an impact on water and food quality in some specific areas which directly implicates human health. Air temperature and precipitation intensity are the most important climatic parameters which affect the distribution of disease pathogens. However, other variables such as sea level elevation and wind also provide significant contributions (Patz et al., 2003; WHO, 2005). Detail of factors and their impact on infectious diseases and its distribution are listed in **Table 6**.

Table 6. Climatic factor and its impact towards disease carrying vectors (Patz et al., 2003; P. Polgreen and E. Polgreen, 2017; WHO, 2018;)

Factor	**Impact**
Temperature	Extreme temperatures can be deadly for the disease-causing pathogens. The temperature change can also engender various effects. It possibly influences the spreads of disease-carrying vectors by modify their biting rates or alter the length of the transmission period. In many cases, the disease-carrying vector will react to the temperature changes by changing the geographical distribution or even adapt to the recent temperature.
Precipitation	The increase of precipitation intensity possibly supplements the growth of disease-carrying vector in many ways. It extends the current habitat size and develops the new breeding grounds. It also increases the food supply which affects the growth of the vertebrate reservoirs population. Heavy rainfall, for example, may cause flooding and at the same time decrease the population of the disease-carrying vector. However, it may force insect and rodent vectors to seek refuge in houses and increases the likelihood of contact with the human.

Humidity	Humidity influences the transmission of insect disease-carrying vector. Mosquito and ticks, for example, can easily vanish on dry conditions. Saturation deficit is an important determinative factor in the case of climatic-based diseases like dengue fever and Lyme disease.
Wind direction and speed	A research performed by Endo and Eltahir (2018) found that wind predisposes the behavior of Anopheles mosquitoes and hence malaria contagion by influences the waves, advection of mosquitoes and Carbon dioxide.
Particulate matter (PM) < 10 µm and < 2.5 µm	PM is inhalable and respirable particles contain sulfate, nitrates, ammonia, sodium chloride, black carbon, mineral dust, and water. It posses the most dangerous risk to human health by penetrating the lungs and entering the blood system.
Carbon monoxide (CO)	A high level of CO is harmful. It will disrupt the amount of oxygen transported to the bloodstream of critical organs.
Nitrogen dioxide (NO_2)	NO2 mainly produced through power generation, industrial as well as traffic activities. Plenty of researches has discovered that independently, it can exacerbate bronchitis, asthma and respiratory infections symptoms, as well as weaken lung function. NO2 possibly also inflicts cardiovascular and respiratory diseases.
Ozone (O_3)	O3 produced as the result of CO, methane or other volatile organic compounds oxidation in the presence of nitrogen oxides and sunlight. O3 conduce inhalation problems, asthma, damage to lung function and respiratory diseases.
Sulphur dioxide (SO_2)	Exposure to SO2 may cause a problem in the respiratory system and lungs. Inflammation on the respiratory system caused by it can even lead to asthma and bronchitis, and also increase the risk of infection.

Water level	Rise of water level caused by climate change may decrease or even eliminate the breeding habitat of salt-marsh mosquitos. Birds and the mammalian host may also be extinct which eliminate the endemic virus. The inland intrusion of salty water can transform the habitats in the freshwater area become the salt-marsh area, displacing the pathogenic agents which occupy the area.

Many research and investigations related to climate change and infectious disease place the human population at great risk. Weather change can close the incubation period of most infectious diseases. Heavy rainfall may increase the vector-borne possibility to spread the diseases while a longer medium temperatures season possibly increases the distribution of the vector-borne itself. During higher temperatures, vectors and pathogens can be easily infectious and quicker in spreading the virus (P. Polgreen and E. Polgreen, 2017).

Despite the availability of many reports related to climatic-based factors for infectious diseases, future work for this field is still required. P. Polgreen and E. Polgreen (2017) listed three major barriers which obstruct the work in this specific area: limited amount of infectious diseases data with specific geographic and time information, requisition of collaboration between infectious disease and computationally-oriented investigators from different fields and the insufficient data from climate-based disease investigations to support this data-science oriented field of study.

2.2.2 Data Warehouse for Disease Surveillance System

During recent years, technology plays an important role to increase the availability of healthcare data. Clinical records are generated through automatic instruments such as mobile applications, satellites, radio-frequency sensors, and various types of wearable devices. Consequently, the healthcare industry can produce a big volume of electronic data. **Fig. 5** depicts the estimated growth of healthcare data from 2013 to 2020.

These data are not only recorded from patient encounters or disease occurence. They can also contain healthcare information although the related person was not looking for medical services. Google, for instance, utilize the internet search query, recent analysis of Wikipedia, and Twitter feeds to track the influenza epidemics and outbreaks. This provides an excellent transformation which, unfortunately, also comes with important drawbacks (Simonsen et al., 2016; Fang et al., 2016).

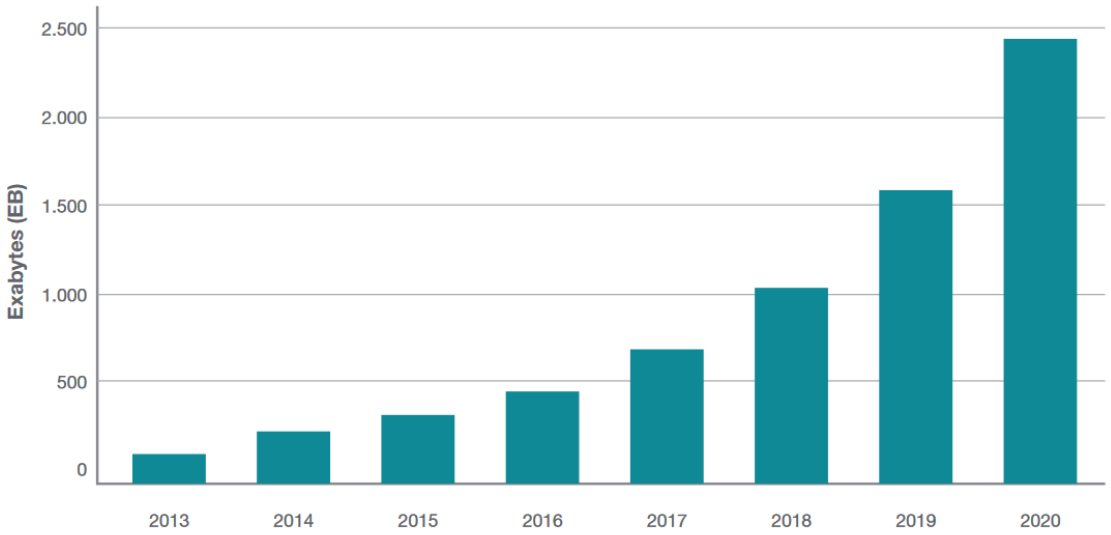

Fig. 5. Growth in healthcare data (Accenture, 2018)

During 2013, International Data Corporation (IDC) estimated that the total volume of global healthcare data was almost 153 exabytes, where one exabyte is equal to one billion gigabytes. This amount is expected to reach 2314 exabytes in 2020, equates to an increase of at least 48% annually (Standford Medicine, 2017). Greenwell and Salentine (2018) define the twelve main data sources in the healthcare area as listed in **Table 7** below.

Table 7. Main sources of healthcare data (Greenwell and Salentine, 2018)

Type of Data	**Data Type**	**Analysis**	**Disaggregation**
Individual records	Morbidity and health conditions; service interventions	Patient or client	Sociodemographic characteristics

Health infrastructure information system	Infrastructure and amenities; types of services; equipment	Facility	Geography; type of facility, management and other
Human resources information system	Health occupation	Health worker	Sociodemographic characteristics
Logistic management information system	Essential medicines and commodities	Medicine or commodity	Geography, type of facility
Financial management information system	Budget estimates; revenue and expenditures	Budget item	(National level)
Health facility assesment	Health resources invetory	Facility	Geography, type of facility
Population census	Population estimates and projections	Person	Sociodemographic characteristics
Population-based survey	Risk factors; knowledge, attitude and practices; coverage of services	Person or household	Sociodemographic characteristics; socioeconomic stratifiers
Civil registry vital statistical system	Births; deaths; stillbirths; causes of death	Person	Sociodemographic characteristics
Public health surveillance system	Reportable conditions; potential public health threats	Disease or event	Geography, other
Collective intervention records	Community (not clinical) interventions	Community	Geography, other
Health accounts	Health financers; health providers; healthcare services or resources consumed	Health expenditure	(National level)

Standford Medicine (2017) states that data connects human day-to-day lives and behavior to tangible health outcomes in three primary areas. The first area is wearable devices such as pedometers and heart rate monitors which continuously collect the patient healthcare data. The second area is direct-to-consumer testing which includes genetic tests and access to online research. The third and last area is medical informal website. The global sale of wearable devices has reached 274 million USD in 2016. Fitness bands are the best-selling type of wearable device among other technologies in this category.

Healthcare data generated through big data streams is not only massive in amount. It also complicated in structure and high at speed. According to Fang et al. (2016), there are five factors which trigger the breakdown of the traditional system in handling the big data. The first factor is data variety in terms of structured and unstructured, for example, handwritten doctor notes, medical records, medical diagnostic images, computed tomography, and radiographic films. The second factor is the heterogeneity and complexity of big data in the healthcare informatics area. The third factor is an impediment to record and analyze these big and diverse data. The fourth factor is the limitation of storage capacity, computation as well as processing power. The fifth factor is the need to improve medical affairs in terms of service quality, use of confidential data and healthcare cost reduction. These problems are simply cannot be solved by the traditional system.

Computerization is widely used to improve the efficiency of healthcare. It mainly serves to record and retrieve information for health providers and patient encounters. The convergence of different areas in healthcare has also promoted the interest of utilizing healthcare data to control and regenerate the quality of healthcare services using data warehouse technology. Many professionals in this domain have recognized that using electronic health data can also beneficial for monitoring infectious diseases and reporting it to the public health authorities. The data warehouse can also help for other purposes like sorting the hospital rooms intended for patients on isolation precautions, control the occurrence of the antimicrobial-resistant organism and calculating trends in antimicrobial use. Following the increase of infectious control departments, the development of data warehouses will also increase the productivity of workers by eliminating the time-consuming and repetitive tasks using an automatic system (Trick, 2008).

Sanders et al. (2017) mentioned that the implementation of data warehouses in healthcare simplifies management reporting and makes it more efficient in three different ways. Firstly, the data warehouse enables an effective and scalable reporting process. It integrates data from disparate sources to elevate the analysis process in a better way. Secondly, the data warehouse guarantees data consistency to be trusted. The data warehouse establishes a single source of truth where everyone can rely on its accuracy to drive critical decisions. Thirdly, the data warehouse enables meaningful and targeted quality improvement. The data warehouse is capable to drive consistent insight, better collaboration and a more streamed process across different departments among healthcare organizations.

Grob and Hartzband (2008) classified the usefulness of the data warehouse in the healthcare area into four categories. On the patient level, take a case of a patient with a chronic condition as an example. Knowledge of their personal medical history, as well as the respond of patients with common traits to this specific medication, will benefit to guide the course of treatments. On the population level, the data warehouse is the pillar of screening initiatives and preventive care measurements. The patient population view with the time and geographic dimension can be obtained by integrating the data warehouse to the geographical information system. For the healthcare provider, the data warehouse can be used as an explorative tool to find out the opportunities of improving the services and outcomes, promote the collegial competition among different organizations considering the data and information transparency and lastly, help to meet the objectives of pay-for-performance initiatives. For healthcare organizations, utilization of data warehouse with appropriate analytics tools will provide BI to gain information about the operation of the organization to improve the decision-making process, develop benchmark do determine the performance of individual centers, plan for operational needs and financial modeling as well as demonstrating performance information in reimbursement level.

Data warehouses in healthcare consist of three levels of data granularity. It covers the coarse-grained data for general report, up to the detail event data such as hospital discharged. Illustration of these three levels of data granularity is shown in **Fig. 6**. To derive different health indicators, all these data can be synthesized with demographic, economic, and marketing data (Berndt et al., 2001).

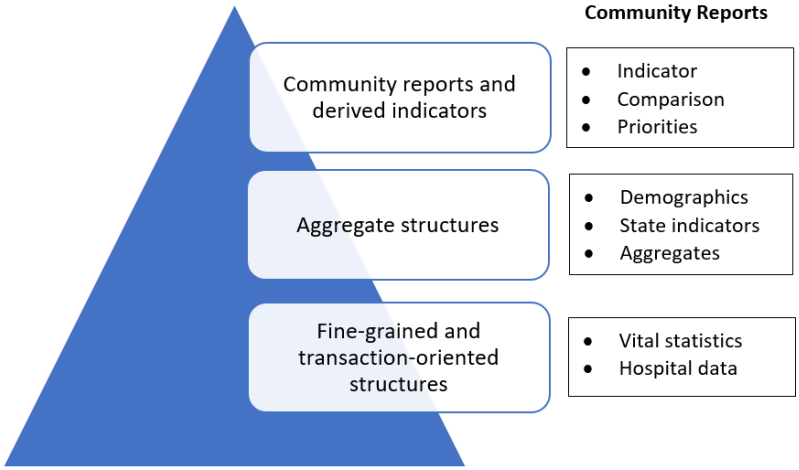

Fig. 6. Three level of data warehous eaggregation (Berndt et al., 2001)

As seen on **Fig.** 6, on top of the pyramid, tables with the highest aggregated data are used to generate the report. These tables are fast and responsive for gaining the data using browsing tools and serves a foundation for basic internet access. In the middle of the pyramid, the aggregate level provides dimensional capabilities including roll-up and drill-down operation at different levels for analysis. On the bottom part, the design retains very fine-grained and event-level data. The facts and dimensions are presented only for analysis and reporting (Berndt et al., 2001). An example of data warehouse architecture for healthcare is illustrated in **Fig. 7**.

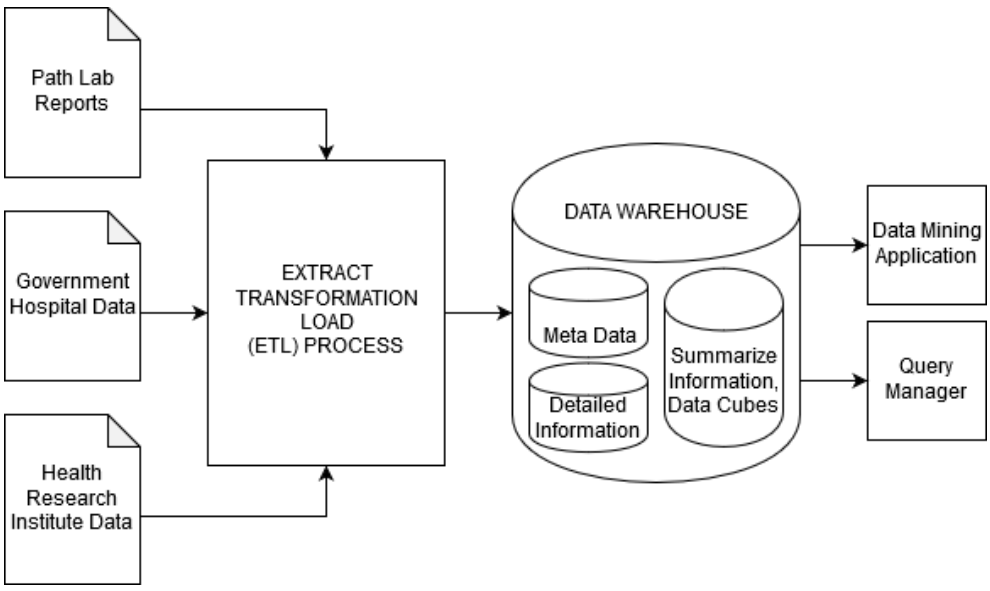

Fig. 7. Example of healthcare data warehouse for healthcare (Khan and Hoque, 2015)

As seen in **Fig. 7**, data are retrieved from different public and private sources systems such as laboratory reports and hospital data to be transferred to the next component for the ETL process. Once it passes the ETL process, data will be retrieved into the data warehouse layer for OLAP queries and mining operations.

A well-design data warehouse for healthcare should possess five characteristics. First, it should depend on multiple sources. The data warehouse must be composed of a minimum of two source systems, be it transactions or operational systems. In large enterprises, it is common that their data warehouses are populated from fifty different sources, both internal and external. Second, it is specifically designed to support cross-organizational analysis. Data warehouses must be qualified to be used for the analysis process ao support the business process. Third, it produces trends-metric-and-reports. Data warehouse produces output characterized by metrics and reports that helps to recognize the trends and hidden relationship in business processes. The fourth and fifth characteristics are large and historical. Data warehouses generally cover billions of records equal to hundreds of terabytes, collected through many years, from five up to 30 years worth (Sanders et al., 2017).

3 DESIGN METHODOLOGY

Although there are many literature reviews discuss software development, only a few research and scientific works have been devoted to data warehouse development and mainly produced based on experience from the real-case project. The scientific community proposed different approaches that generally targeting the specific conceptual model; however, those approaches are too sophisticated to be applied in the real-world environment. Consequently, there still a lack of methodological framework as guidance for data warehouse developers during the development process.

Traditionally, the data warehouse design methodology consists of four closely related yet not necessarily strictly sequential phases. The data warehouse is considered as a particular type of database, used for the analytical process. Therefore, it acceptable to apply the design methodology of the traditional database into the data warehouse design. However, there are several significant differences, accounted for their different natures (Vaisman and Zimányi, 2014). The whole phase of the data warehouse design methodology is illustrated in **Fig. 8**.

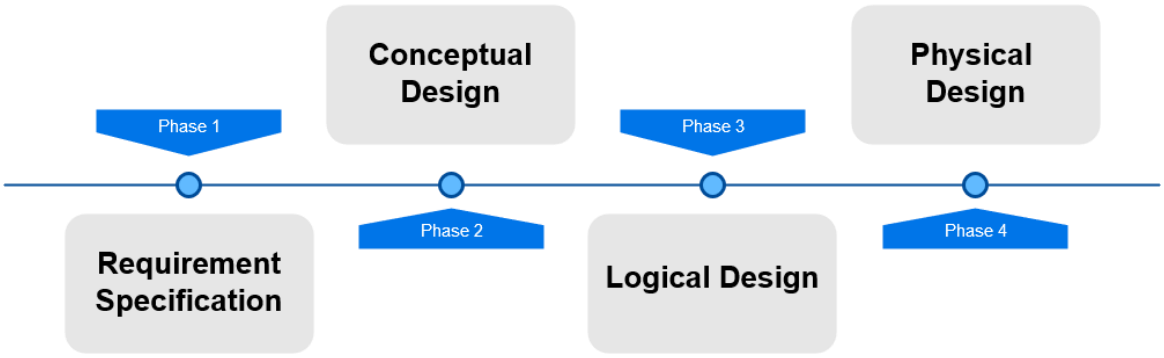

Fig. 8. Phases in data warehouse design

These four phases can be applied both for the bottom-up as well as top-down approaches. However, the distinctive characteristic will be given to the requirement specifications and conceptual design. There are three sub-categories which affect these two phases: analysis-driven approach, source-driven approach, and analysis/source-driven approach.

The participation of users from different levels is mandatory when it comes to the data warehouse design with the analysis-driven approach. Key users need to be identified first, with consideration of the following issues: users must understand the overall business goals instead of personal perception, users must be cooperative in sense of not being dominated and tempered, users must be active and possess adequate knowledge about data warehouse and OLAP system. The analysis-driven approach requires the development team with strong competence in leadership and communication in addition to high technical skills.

Contras with the above description, user participation is not substantial in the data warehouse design with the source-driven approach. Instead, this approach requires an analysis to underline the source system and obtain the warehouse schema during the initial phase. Users come from professional or administrative levels with the main role to confirm the data structure, facts, and measurements as the basis to develop the multidimensional schemas. The source-driven approach demands highly qualified designers with experiences in the technical and business domain.

The analysis/source-drive approach is a combination of the analysis-driven and source-driven approaches. It considers both the business needs as delivered by the key users as well as the availability and accessibility of the source systems where the data come from. The development teams for this approach are a consolidation of teams recommended for an analysis-driven and source-driven approach.

3.1 Requirements Specification

Requirements specification is the first phase in the data warehouse design methodology. Therefore, it entails a significant problem if it is done incorrectly or incompletely (Golfareli and Rizzi, 2009). According to Kimball and Ross (2013), there are two primary techniques to gather the requirement specification: interviews and facilitated sessions. Interviews encourage users to actively participate and generally resulted in a detailed list of the requirement specification. However, it can also be difficult and fruitless as different

languages used by designers and users. The facilitated session involved more participants, led by a facilitator who responsible for setting up a common language for all the interviewees. It is useful for creative brainstorming; however, it is more difficult to be scheduled and required more works.

Different means were given to requirements specifications. The analysis-driven approach requires analysis of the user needs in its early stage to clarify the goals and needs of the organization in implementing the data warehouse. The following steps are taken afterward: determining the analysis needs and documenting the result. Source-driven source hinges on available data on the source system. This information will be used to identify the multidimensional schemas to be implemented. These databases will be analyzed to figure out the representing elements of facts, dimensions, hierarchies, and measures. These findings will be used as a foundation for the first conceptual schema. Therefore, identifying the source systems is indicated as the first step. The next steps are applying the derivation process and documenting the result. The analysis/source-driven approach to requirement specification consolidates steps from both approaches simultaneously to achieve the most optimal design solution. The sequence steps are defining the analysis needs and initial elements for the multidimensional schema (Vaisman and Zimányi, 2014).

3.2 Conceptual Design

A well-executed requirement specification phase should be able to deliver a clear and concise schema. This schema will be used as the foundation to build the data warehouse's initial concept. There is no universal standard of models for conceptual design in the data warehouse. The entity-relationship model is widely used. However, Kimball and Ross (2013) stated that it cannot be used since the EDW has different degrees of normalization. They propose the star schemas or star joins. Nevertheless, Inmon (2002) mentioned that star joins can only be used for data marts.

The conceptual model for each project depends on the specific needs of the project itself. This research will adopt the MultiDim model introduced by Vaisman and Zimányi (2014).

It represents the conceptual level of data warehouse elements and application of Online Analytical Processing (OLAP) which are dimensions, hierarchies, and facts with all the associated measures. **Fig. 9** below show the graphical notation of a MultiDim model for representing the conceptual design of a data warehouse.

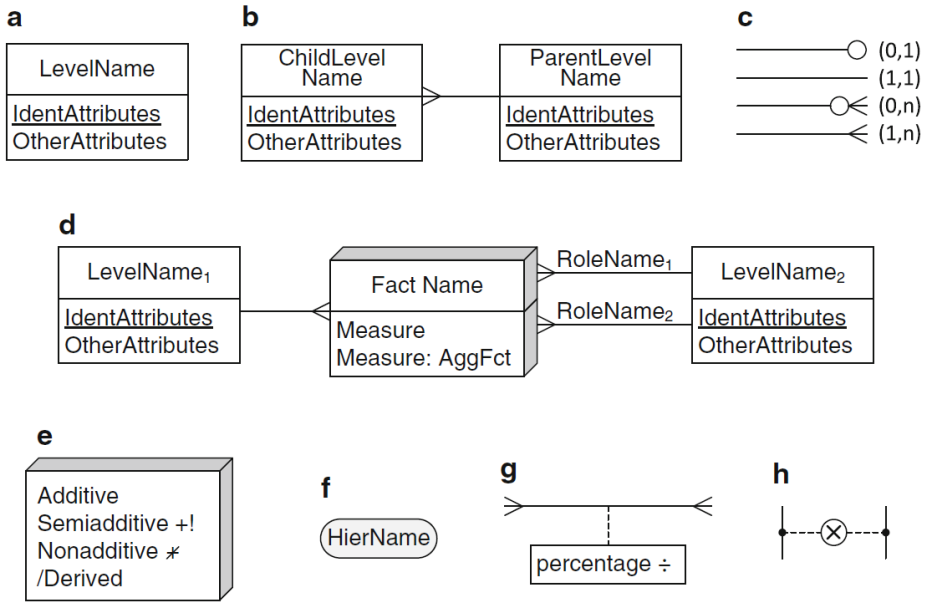

Fig. 9. Notation of MultiDim model: (a) level, (b) hierarcy, (c) cardinalities, (d) fact with measures and associates level, (e) types of measures, (f) hierarcy name, (g) hierarcy attributes, (h) exclusive relationships (Vaisman and Zimányi, 2014)

The key elements of the MultiDim model are the schema, dimension, level, and fact. Schema is the illustration of the data warehouse, consists of a set of dimensions and facts which logically related to each other. Dimension is a reference for a measurable event, composed by either one level or minimum one hierarchy which instead, consists of a set of levels. There is no graphical notation to represent the dimension as it is depicted by the elements. The level is a description of a general concept with similar characteristics from the application view. It consists of a set of attributes to represent the characteristics of each member and identifier. The fact used to related different levels. It may contain measures, the numerical data that aggregated along dimension during roll-up operations. There are three classifications of measures: additive, semi-additive, and non-additive. Hierarchy comprised of several related levels. The lower level termed as the child and the higher level termed as the parent.

Similar to the foregoing phase, different steps were given to the conceptual design. Analysis-driven conceptual design is an iterative process consists of the development of the initial schema, verification of data availability in the source systems and data mapping in the schema and the sources. Modification of schema required in case of missing data item. The source-driven conceptual design consists of defining the first schema, validating the first proposed schema with users and defining the final schema and mappings. Analysis/source-driven approach includes three sequential steps as well: define, correct and mapping the conceptual schema (Vaisman and Zimányi, 2014).

3.3 Logical Design

Schema produced on conceptual design will be converted to the logical schema for the data warehouse development during logical design. Three approaches were used to implement the multidimensional model in this phase, depends on the data cube storage: Relational OLAP (ROLAP), Multidimensional OLAP (MOLAP) and Hybrid OLAP (HOLAP). In ROLAP, data are stored in the relational database. Aggregates are also precomputed in relational tables to increase the performance. In MOLAP, data cubes are reserved in a multidimensional array together with the hasting and indexing technique to make the operation efficiently implemented. The data management performed by the multidimensional engine to reduce storage space. HOLAP systems take advantage of the storage capacity from ROLAP approach and the processing capacity of MOLAP. It may storage large volumes of detail data in the relational database but the aggregations are preserved separately in the MOLAP store. (Golfareli and Rizzi, 2009; Vaisman and Zimányi, 2014).

Multidimensional model on logical design commonly represented by the star schema, also known as star join and snowflakes schema. Example of star schema and snowflakes schema are shown in **Fig. 10** below.

Star schema is a relational schema composed of one central fact table and a set of the dimension table. The fact table contained foreign keys of all associated dimension tables

while the referential integrity constraints are specified in between. Normally, the dimension tables are not normalized. It is possibly contained redundant data especially if there are hierarchies. Only the fact table that is usually normalized, the union with foreign keys functionally determines all the measures without dependency among key attributes. Snowflake schema on the opposite eludes the redundancy with normalization of dimension tables. The snowflake schema is obtained by breaking down the dimension tables from star schema into smaller tables to remove the dependencies of transitive functional.

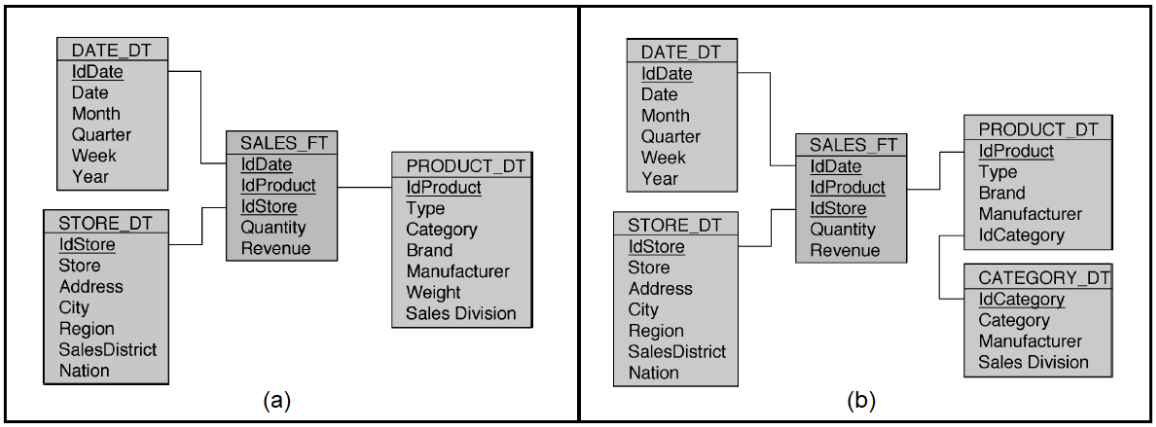

Fig. 10. Example of (a) star schema and (b) snowflake schema (Rizzi, 2008)

The logical design consists of two consecutive steps: define the logical schema and define the ETL process. To deliver the logical schema, it is important to apply the general rules of mapping to the multidimensional schema from conceptual design. ETL refers to the extraction function for collecting data from the sources, transformation function for adjusting the data format and loading function for entering the transformed data into the data warehouse. The preliminary sequence of the ETL process is needed to ensure that all data will be transformed to fulfill the specific standardization without overriding its consistency (Golfareli and Rizzi, 2009; Vaisman and Zimányi, 2014).

3.4 Physical Design

Physical design is the process of converting the proposed schema into actual database structures by mapping the entities to the table, relationships to the foreign key constraints,

attributes to the columns, primary unique identifiers to the primary key constraints and unique identifiers to the unique key constraints. During this phase, the logical schema from previous phase will be converted into a tool-dependent physical structure; furthermore, model for the data warehouse will be define, consist of entities, attributes and relationship. This is important in order to ensure sufficient time to response the query. As the logical design and physical design are closely interdependent, it is preferable to execute both phases together to achieve the best result. A comparison of logical and physical design during data warehouse development is shown in **Fig. 11** as follows.

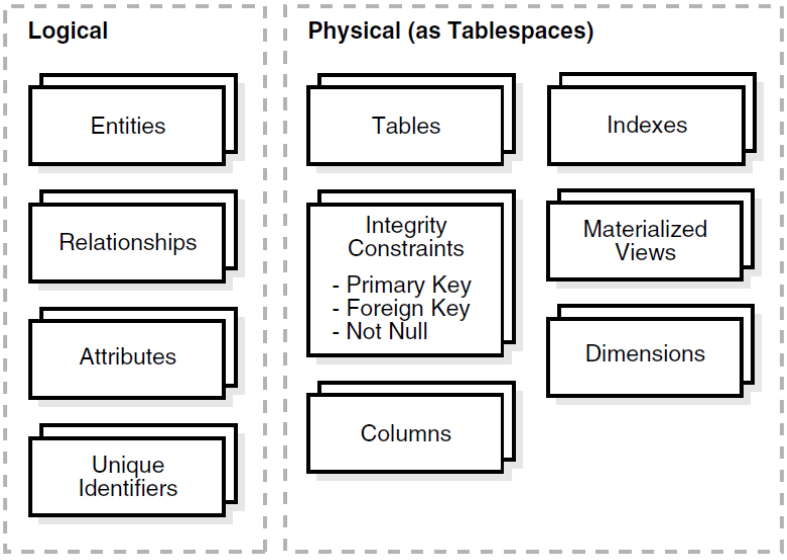

Fig. 11. Logical vs physical design (Lane, 2013)

Logical design is executed by drawing with a pen and paper or design with a data warehouse builder application. Physical design is creating the database typically with SQL statements. This phase includes gathers all the data from the logical design phase and converts them into a description of the physical database structure. The main driver of data warehouse physical design is query performance and database maintenance (Lane, 2013).

Implementation of these three techniques will escalate the data warehouse performance: materialized views, indexing, and partitioning. The materialized view is the best option to generate good query performance in OLAP. It uses the physical storage in the database to improve the performance of the query by precalculating the cost of operation such as joins and aggregation. Indexing is the best option to generate query performance in the data

warehouse. During the physical design phase, it is required to define which kinds of indexes that will be used over which attributes. B-tree and hashes indexes are typically used for the database management system, while bitmap and join indexes are generally used in the data warehouse. Partitioning or fragmentation will separate the contents of relation into different files on range values of the attributes. It consists of dividing the table into smaller datasets and provide better support for managing a large volume of data (Vaisman and Zimányi, 2014).

Similar to the logical design phase, the physical design phase also consists of two steps: one step related to the implementation of data warehouse schema while the other step is for the ETL process. A well-develop physical design will produce a data warehouse system with the ability to manage big data, automatically update the current data warehouse with new data extracted from the sources system, execute sophisticated operations include joins of different tables and aggregating various data items. All of these function depends on the method of the storage, indexing, partitioning, execution of parallel query execution, the function of aggregation and view materialization of the developed data warehouse (Lane, 2005; Vaisman and Zimányi, 2014).

4 DATA WAREHOUSE DESIGN

The data warehouse design in this chapter will follow the theoretical framework described in the previous chapter. Additionally, two research papers with similar objectives and limitations were also taken into consideration: Yun et al. (2011), and Ivančević et al. (2013). Considering the characteristics and availability of the resources, the data warehouse architecture of this thesis work will adopt the bottom-up architecture proposed by Kimball with an analysis/source-driven approach specifically applied to the requirement specification and conceptual design phase.

4.1 Requirement Specification

Following the framework presented during the foregoing chapter, the requirement specification of this proposed data warehouse is divided into three sequential steps: identify users and analysis needs, identify the source systems and apply the derivation process.

4.1.1 Identify the Users and Analysis Needs

Intended users for this proposed data warehouse were categorized into three groups. First, medical experts in healthcare organizations who deal with infectious disease and epidemiological issues. Instead of using other solutions intended for generic statistical analysis, they can rely on data provided on this system as it is purposely designed for the analysis process in the area. By having this domain-specific system, productivity can be improved. Second, scientists and academia whose topic of research is related to infectious diseases and epidemiology. They will be able to develop, test and improve the model by utilizing data provided on this system. Additionally, the proposed data warehouse can also be used by users who are not medical experts or scientists but interest to delve the latest infectious disease trends, forecasts or results of the specific analysis in this area.

The purposes of this proposed data warehouse were also listed into three main points. First, to design an integrated repository consisted of climate parameters (weather observation, air quality observation, and marine observation) and infectious disease case register in Finland. Second, to support forecasting of the currently listed infectious disease as listed in **Appendix 1** as well as the growth of the infectious diseases in a specific area. This can be done by examining the correlation between climate parameters and infectious disease occurrences in that particular area during a specific period. Thirdly, to evaluate the potential increase of infectious pathogens in a specific area as the impact of climate change.

4.1.2 Identify the Source Systems

There are four main datasets used for this data warehouse development. The first dataset is the infectious disease register data, taken from the Finnish Institute for Health and Welfare. This dataset, however, only available in Finnish or Swedish. Fortunately, there is not much translation needed for this as mainly data used from this source are numerical. The second dataset is weather observation data, the third dataset is air quality observation and the fourth dataset is marine observation data. The second, third and fourth datasets were taken from the Finnish Meteorological Institute. All data are available in excel and CSV format, free of charge for public use from their official website. The description of the main data used for this proposed data warehouse were listed in **Table 8**.

Table 8. Description of data used on proposed data warehouse

Datasets		**Description**
Infectious Disease Register	Reporting group	Disease name and disease category
	Time	Period of infectious disease registration, from weekly to annual basis. Data are available from the year of 1995 onwards
	Area	The location where the specific disease occurred, categorized by region to district level

	agegroup	Cases were divided into 5 groups of age with 5 years interval, from 0 to over 75 years old
	sex	Cases were divided into 3 sex categories: woman, man and not available (n/a)
	Indicator	This proposed data warehouse will use case as the default parameter
Weather Observation	Station	Observation station where the data recorded
	Year	Year of the observation data recorded
	m	Month of the observation data recorded
	d	Day of the observation data recorded
	Time	Hour of the observation data recorded
	Relative humidity (%)	Comparison of the current humidity present in the air with its highest possibility, presented in Percent (%)
	Precipitation intensity (mm/h)	Rate of precipitation, presented in millimeters per hour (mm/h)
	Air temperature (degC)	Degree of hotness or coldness measured on degree celcius
	Wind direction (deg)	Direction where the wind coming from, measured in degrees, clockwise from due north
	Wind speed (m/s)	Speed of air moving past specific area, measured in meters per second (m/s)
Air Quality Observation	Station	Observation station where the data recorded
	Year	Year of the observation data recorded
	m	Month of the observation data recorded
	d	Day of the observation data recorded
	Time	Hour of the observation data recorded
	Carbon monoxide (ug/m3)	Amount of CO on specific area, measured in micrograms per cubic meter (ug/m3)
	Nitrogen dioxide (ug/m3)	Amount of NO_2 on specific area, measured in micrograms per cubic meter (ug/m3)

	Ozone (ug/m3)	Amount of O_3 on specific area, measured in micrograms per cubic meter (ug/m3)
	Particulate matter < 10 µm (ug/m3)	Amount of PM less than 10 micron on specific area, measured in micrograms per cubic meter (ug/m3)
	Particulate matter < 2.5 µm (ug/m3)	Amount of PM less 2.5 micron on specific area, measured in micrograms per cubic meter (ug/m3)
	Sulphur dioxide (ug/m3)	Amount of SO_2 on specific area, measured in micrograms per cubic meter (ug/m3)
Marine Observation	Station	Observation station where the data recorded
	Year	Year of the observation data recorded
	m	Month of the observation data recorded
	d	Day of the observation data recorded
	Time	Hour of the observation data recorded
	Water level (mm)	Elevation of free surface on coastal area, measured in millimeters (mm)

The weather, air quality, and marine observation data can only be downloaded separately from each observation station during a specific period. Therefore, the station, Year, m, d and time columns were added on each table. While checking the data availability, a problem was discovered: the weather, air quality, and marine observation data are not available from all observation stations. Moreover, it is also possible that data may contain columns with empty value. To ensure the data consistency, all these empty values will be removed on the ETL process.

It is possible to include all columns on infectious disease register data into one single file. However, the clarity of data will be indicated clearer by applying lesser variables and filters, for example sorting the total case of the specific disease occurred in a specific area, during a specific time. Moreover, considering the climate data using on this proposed data warehouse as well as the data availability problem as mentioned above, it is more reasonable to narrow down the observation into the particular area as well as time and period.

4.1.3 Apply Derivation Process

Based on the source systems and characteristics of data to be stored, an entity-relationship schema was constructed as depicted in **Fig. 12**.

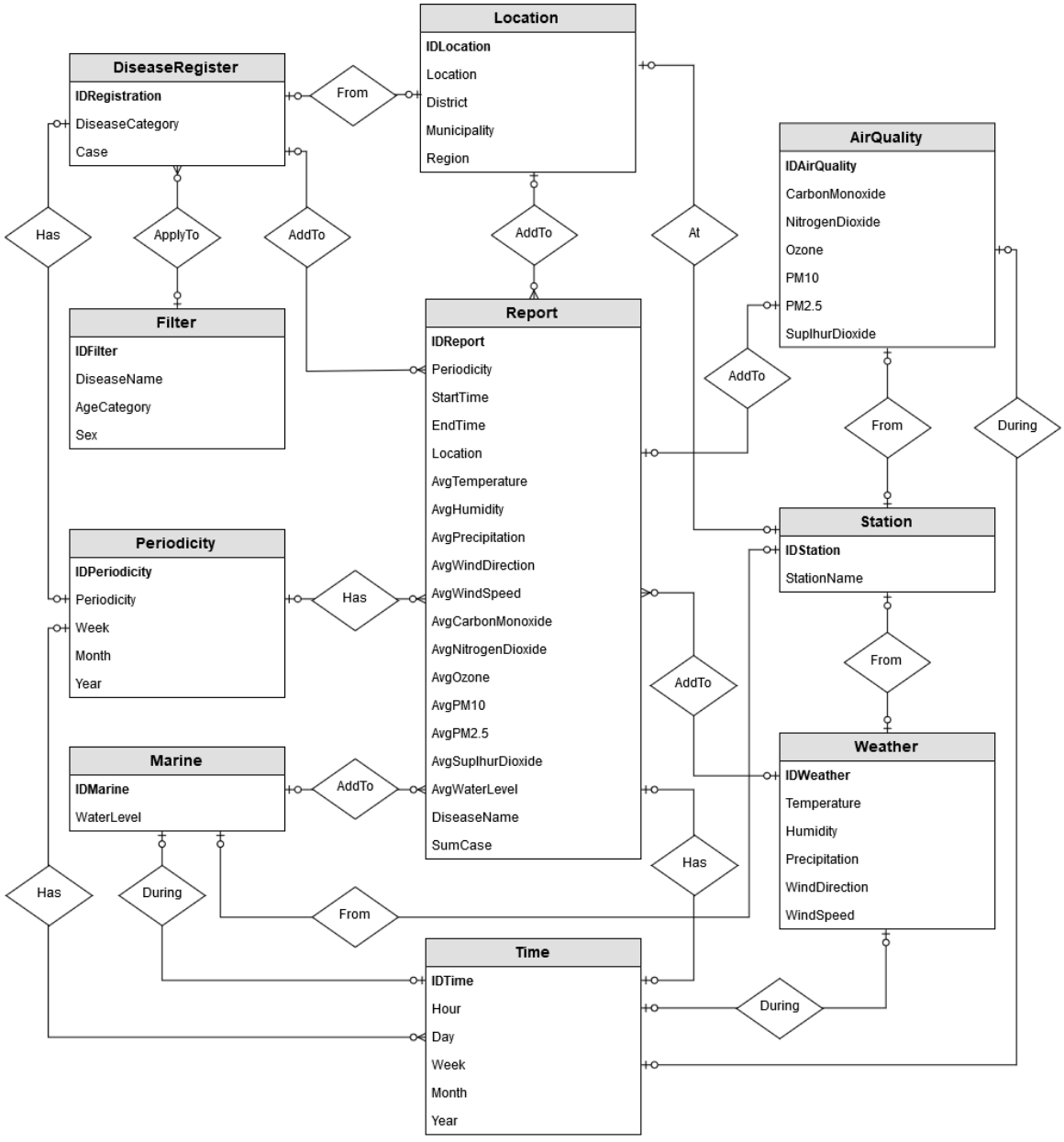

Fig. 12. Entity-relation schema for proposed data warehouse

As seen on the relational schema, it is possible to recognize the many-to-many relationship from `Report` with attributes representing numerical data. Therefore, it can be selected as a fact candidate in the multidimensional schema. The summary of the obtained

multidimensional elements are listed in **Table 9**.

Table 9. List of obtained multidimensional elements for proposed data warehouse

Fact	Measures	Dimension	Relation	Hierarchies
Report	AvgTemperature AvgHumidity AvgPrecipitation AvgWindDirection AvgWindSpeed AvgCarbonMonoxide AvgNitrogenDioxide AvgOzone AvgPM10 AvgPM2.5 AvgSulphurDioxide AvgWaterLevel SumCase	Periodicity	1:n	**Categories** PeriodicityName → PeriodicityCategory
		StartTime	1:n	**Time** Hour → Date → Month → Year
		EndTime	1:n	**Time** Hour → Date → Month → Year
		Location	1:n	**Geography** District → Municipality → Region

The candidate measures for Report are AvgTemperature, AvgHumidity, AvgPrecipitation, AvgWindDirection, AvgWindSpeed, AvgCarbonMonoxide, AvgNitrogenDioxide, AvgOzone, AvgPM10, AvgPM2.5, AvgSulphurDioxide, AvgWaterLevel and SumCase. Furthemore, the candidate dimensions are Periodicity, StartTime, EndTime and Location.

4.2 Conceptual Design

Similar to the requirement specification phase, conceptual design for the proposed data warehouse also performed according to the theoretical framework explained in chapter three. However, there are some modifications applied due to the limitation and nature of this work. Adopting the analysis/source-driven approach, this phase consists of two sequential steps: develop the conceptual schema and specify mappings.

4.2.1 Develop the Conceptual Schema

Based on the requirement analysis on the previous phase, the conceptual diagram is constructed as depicted in **Fig. 13**.

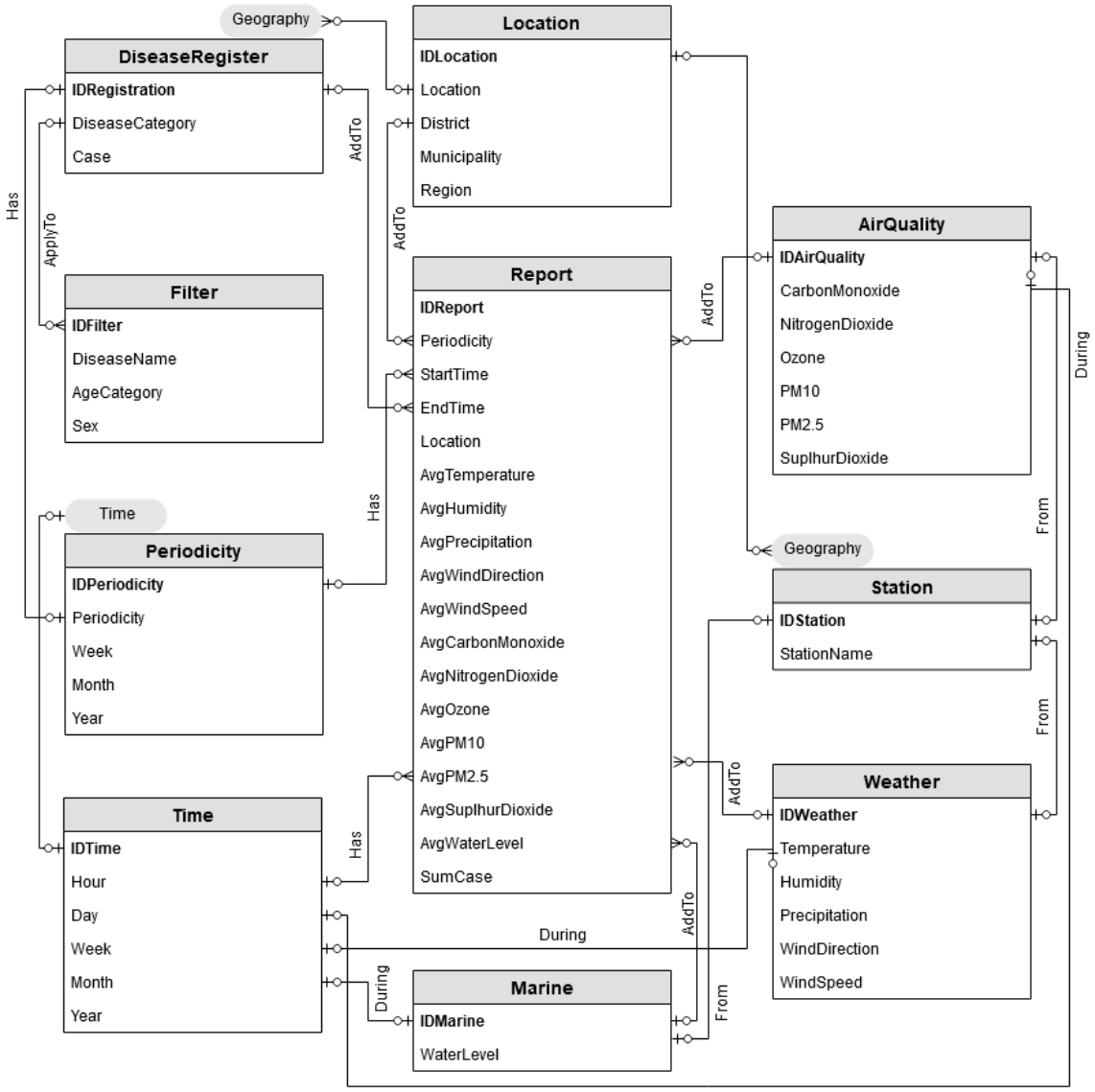

Fig. 13. Conceptual schema for proposed data warehouse

As seen in **Fig. 13**, there is one fact table: `Report` with nine dimension tables: `Periodicity`, `Time`, `Location`, `Weather`, `AirQuality`, `Marine`, `Station`, `DiseaseRegister`, and `Filter`. Each table has one primary key.

4.2.2 Specify the Mappings

After constructed the conceptual schema, the next step is to map the columns on source systems into the proposed schema. As the column may be fetched from different tables or even sources, it is also required to determine the transformation or calculation to obtain the source column into the target column. **Table 10** indicated data transformation conducted to the dataset retrieved from its source systems.

Table 10. Data Transformation conducted to dataset retrieved from the source systems

Source System		**Data Warehouse**		**Transfor -mation**
Table	**Column**	**Level**	**Attribute**	
Infectious Disease Register	Reporting group	DiseaseRegister	DiseaseCategory	✓
	Time	Periodicity	Periodicity	✓
	Area	Location	District	✓
	agegroup	Filter	AgeGroup	✓
	sex	Filter	Sex	✓
	Indicator	DiseaseRegister	Case	✓
Weather Observation	Year	Time	Year	✓
	m	Time	Month	✓
	d	Time	Day	✓
	Time	Time	Hour	✓
	Relative humidity (%)	Weather	Humidity	✓
	Precipitation intensity (mm/h)	Weather	Precipitation	✓
	Air temperature (degC)	Weather	Temperature	✓
	Wind direction (deg)	Weather	WindDirection	✓
	Wind speed (m/s)	Weather	WindSpeed	✓
Air Quality Observation	Year	Time	Year	✓
	m	Time	Month	✓

	d	Time	Day	✓
	Time	Time	Hour	✓
	Carbon monoxide (ug/m3)	AirQuality	CarbonMonoxide	✓
	Nitrogen dioxide (ug/m3)	AirQuality	NitrogenDioxide	✓
	Ozone (ug/m3)	AirQuality	Ozone	✓
	Particulate matter < 10 µm (ug/m3)	AirQuality	PM10	✓
	Particulate matter < 2.5 µm (ug/m3)	AirQuality	PM2.5	✓
	Sulphur dioxide (ug/m3)	AirQuality	SulphurDioxide	✓
Marine Observation	Year	Time	Year	✓
	m	Time	Month	✓
	d	Time	Day	✓
	Time	Time	Hour	✓
	Water level (mm)	Marine	WaterLevel	✓

As seen in **Table 10**, the transformation is applied to all columns on the dataset retrieved from the source systems. The transformation procedure including removing the unnecessary column from the original dataset and also to change the name of the column to avoid space as well as fulfilling the consistency requirement of data to ensure the further process become easier.

4.3 Logical Design

Star schema and snowflakes schema are used to represent the logical design model for the proposed data warehouse. Star schema is depicted in **Fig. 14**.

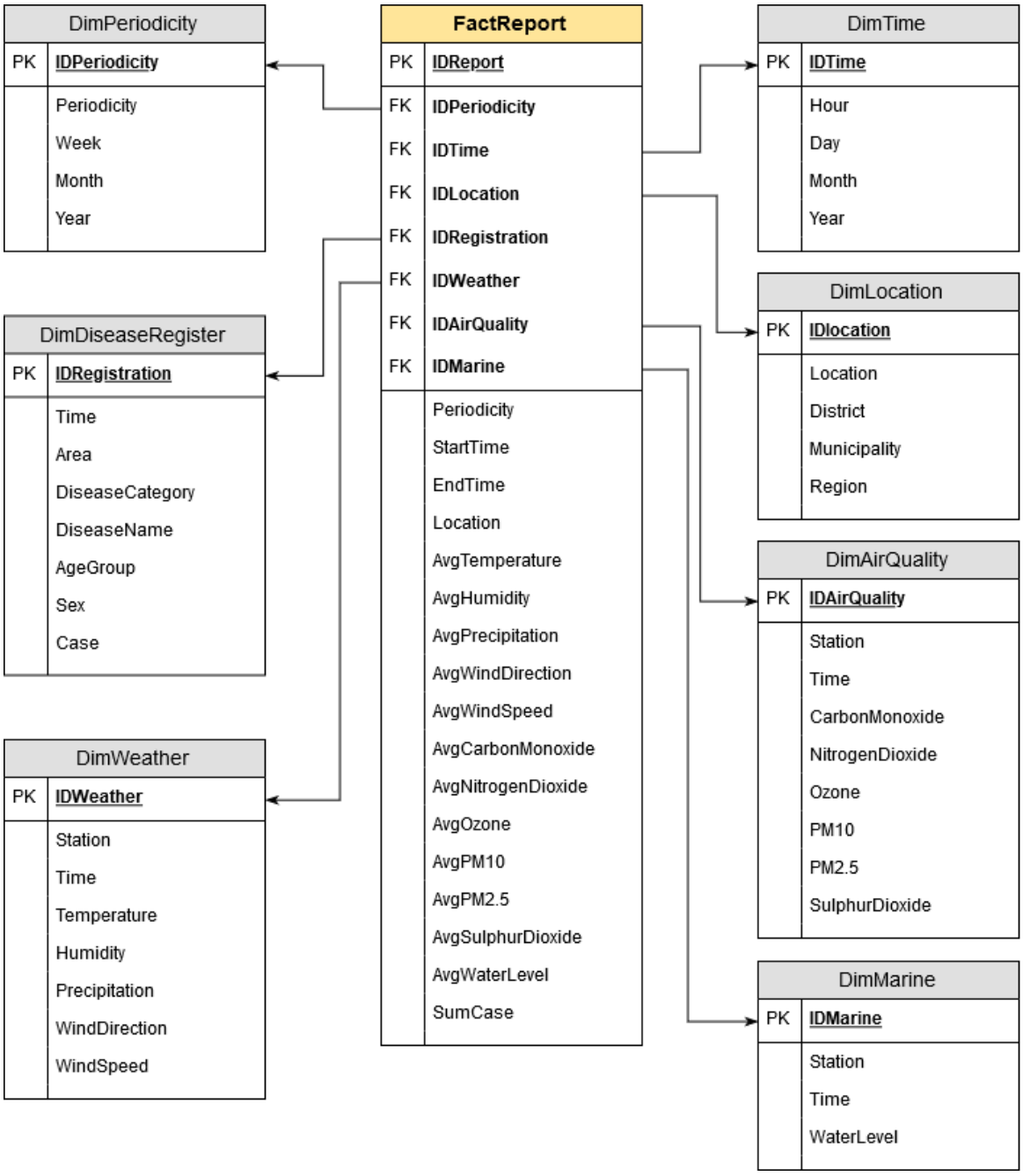

Fig. 14. Star Schema for proposed data warehouse

In the star schema, there is one central fact table: `Report`, which contains the foreign key of all related dimension tables: `IDPeriodicity`, `IDTime`, `IDLocation`, `IDWeather`, `IDAirQuality`, `IDMarine`, and `IDDiseaseRegister`. The fact table contains all measures from multidimensional elements listed in **Table 9** during the requirement specification phase. The snowflake schema is depicted in **Fig. 15**.

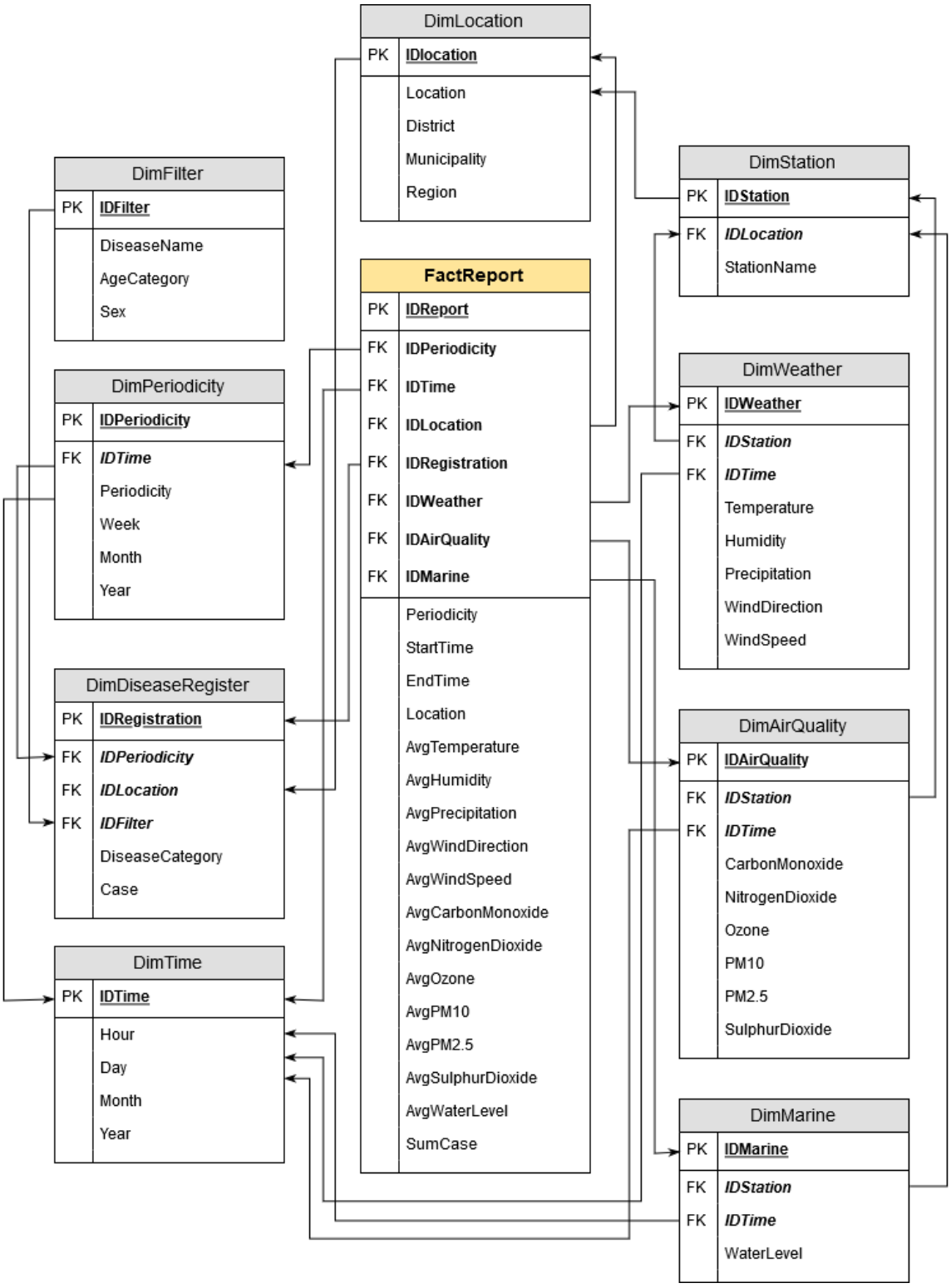

Fig. 15. Snowflakes schema for proposed data warehouse

The representation of logical design with snowflakes schema looks more complex. The fact table contained the same foreign key and measures as the previous schema. However,

two more dimension tables were added to avoid redundancy: `Filter` and `Station`. **Table 11** provide the description of the attribute on the `FactReport` table.

Table 11. Description of attribute on `FactReport` table

Column	Type	Description
`IDReport`	int	Key of the report
`IDPeriodicity`	int	Key of periodicity applied on report
`IDTime`	int	Key of time when the data used on the report retrieved
`IDLocation`	int	Key of location where the data used on the report retrieved
`IDRegistration`	int	Key of infectious disease registration data used on the report
`IDWeather`	Int	Key of weather observation data used on the report
`IDAirQuality`	int	Key of air quality observation data used on the report
`IDMarine`	int	Key of marine observation data used on the report
`Periodicity`	varchar	Periodicity type applied on the report
`StartTime`	datetime	Start time of observation and disease registration
`EndTime`	datetime	End time of observation and disease registration
`Location`	varchar	Location where the data retrieved
`AvgTemperature`	numeric	Average of temperature measurement (degree Celcius)
`AvgHumidity`	numeric	Average of humidity measurement (%)
`AvgPrecipitation`	numeric	Average of precipitation measurement (mm/h)
`AvgWindDirection`	numeric	Average of wind direction measurement (degree)
`AvgWindSpeed`	numeric	Average of wind speed measurement (m/s)
`AvgCarbonMonoxide`	numeric	Average of CO measurement (ug/m3)
`AvgNitrogenDioxide`	numeric	Average of NO_2 measurement (ug/m3)
`AvgOzone`	numeric	Average of O_3 measurement (ug/m3)

AvgPM10	numeric	Average of PM less than 10 µm measurement (ug/m3)
AvgPM2.5	numeric	Average of PM less than 2.5 µm measurement (ug/m3)
AvgSulphurDioxide	numeric	Average of SO_2 measurement (ug/m3)
AvgWaterLevel	numeric	Average of water level measurement (ug/m3)
SumCase	numeric	Total of disease case

Primary key for FactReport is IDReport. This fact table also contained foreign key from all associated dimension tables: IDPeriodicity, IDTime, IDLocation, IDRegistration, IDWeather, IDAirQuality and IDMarine. The measurements are average of all weather, air quality and marine paramenters: AvgTemperature, AvgHumidity, AvgPrecipitation, AvgWindDirection, AvgWindSpeed, AvgCarbonMonoxide, AvgNitrogenDioxide, AvgOzone, AvgPM10, AvgPM2.5, AvgSulphurDioxide, and AvgWaterLevel as well as SumCase to represent total occurrence of disease. The dimensions are Periodicity, StartTime, EndTime, and Location.

4.4 Physical Design

The physical design of this proposed data warehouse is modeled based on snowflakes schema as depicted in **Fig. 15**. The basic architecture for this proposed data warehouse is illustrated in **Fig. 16**. The data warehouse is composed of four principal components that serve specific functions: data sources, data staging, data warehouse, and data presentation. The explanation of each component is presented in chronological order.

As described in the previous section and also illustrated in **Fig. 16**, there are four data sources for the proposed data warehouse: infectious disease register data, weather observation data, air quality observation data and marine observation data. The tables and columns from each dataset used for this proposed data warehouse are listed in **Table 8**.

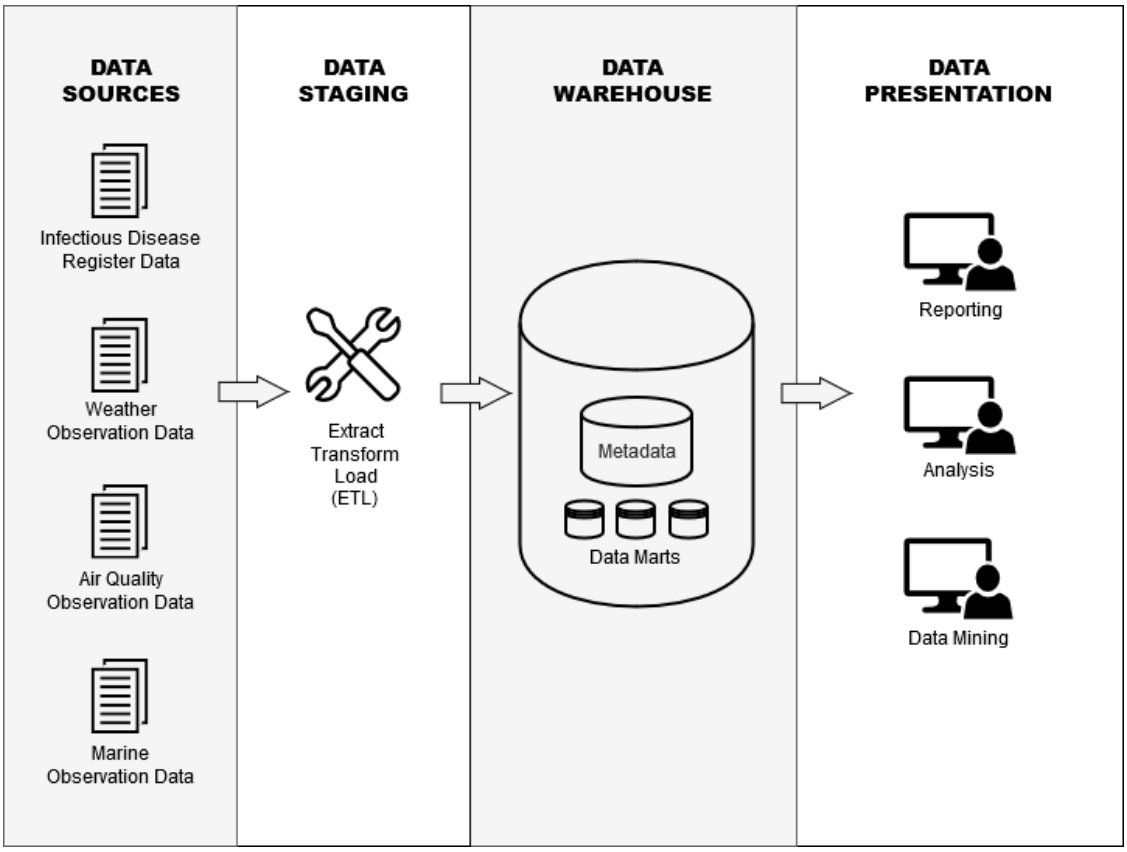

Fig. 16. Architecture for proposed data warehouse

Data staging is the place where the ETL process took place. Data extracted from the data source systems will be modified to satisfy the required format. Once the transformation process is done, data will be loaded to the data warehouse. An example of data extracted from the data source is shown in **Table 12**.

Table 12. Example of dataset extracted from source system

Year	m	d	Time	Time zone	Cloud amount (1/8)	Pressure (msl) (hPa)	Relative humidity (%)	Precipitation intensity (mm/h)	Snow depth (cm)	Air temperature (degC)	Dew-point temperature (degC)	Horizontal visibility (m)
2018	1	1	00:00	UTC	8	1004.3	92	0	6	-1.1	-2.2	38182
2018	1	1	01:00	UTC	8	1003.6	92	0	6	-1.3	-2.5	32229
2018	1	1	02:00	UTC	8	1003	92	0	6	-1.4	-2.6	37839
2018	1	1	03:00	UTC	8	1002.4	91	0	6	-1.4	-2.6	29224
2018	1	1	04:00	UTC	8	1001.4	92	0	6	-1.5	-2.6	18216
2018	1	1	05:00	UTC	8	1000.4	94	0	6	-1.5	-2.4	9213
2018	1	1	06:00	UTC	8	999.9	96	0.8	7	-1.6	-2.2	1207
2018	1	1	07:00	UTC	8	999.5	97	0.3	7	-1.6	-2.1	1754

The transformation process is done by removing unnecessary columns on the original data set. The next process is to fix the mismatch between data extracted from the source

systems and data warehouse schema by implementing a different kind of operation such as reformatting, recalculating, adding the time column, summarizing, merging data from different sources, etc. An example of transformed data to be loaded into the data warehouse is shown in **Table 13**.

Table 13. Example of data to be loaded into proposed data warehouse

Year	Month	Date	Hour	Humidity	Precipitation	Temperature	WindDirection	WindSpeed
2019	1	1	00:00	94	60.3	-0.9	180	5.7
2019	1	1	01:00	95	54.2	-0.9	180	5.7
2019	1	1	02:00	95	57.6	-0.9	180	4.6
2019	1	1	03:00	94	67.2	-0.8	190	5.1
2019	1	1	04:00	95	67.5	-0.7	180	4.6
2019	1	1	05:00	95	64.2	-0.4	170	4.6
2019	1	1	06:00	96	68.3	-0.2	160	4.1
2019	1	1	07:00	96	52.9	-0.3	160	4.1
2019	1	1	08:00	97	60.8	-0.1	160	3.6
2019	1	1	09:00	97	67.4	0	160	4.1
2019	1	1	10:00	97	61.2	-0.1	150	3.6
2019	1	1	11:00	97	69.4	-0.1	150	3.6
2019	1	1	12:00	98	70.9	-0.1	140	3.1

The third component is the data warehouse, where all the transformed data are stored. It includes nine dimension tables and one fact table: `DimPeriodicity`, `DimTime`, `DimLocation`, `DimDiseaseRegister`, `DimFilter`, `DimWeather`, `DimAirQuality`, `DimMarine`, `DimStation`, and `FactReport`. While the description of the `FactReport` table is listed in **Table 11**, descriptions of all the dimension tables are listed in **Table 14**.

Table 14. Description of dimension tables for proposed data warehouse

Table/column	**Type**	**Description**
`DimPeriodicity`	Table	Periodicity applied to the report
`IDPeriodicity`	int	Primary key
`IDTime`	int	Foreign key of DimTime
`Periodictity`	varchar	Weekly, monthly, yearly
`Week`	tinyint	Week of the applied periodicity
`Month`	tinyint	Month of the applied periodicity

Year	int	Year of the applied periodicity
DimTime	table	Time when the data recorded
IDTime	int	Key of time when the data retrieved
Hour	time	Hour of the day
Day	tinyint	Day number
Month	tinyint	Month number
Year	tinyint	Year
DimLocation	table	Location where the data recorded
IDLocation	int	Key of location where the data retrieved
Location	char	Location name
District	char	District name
Municipality	char	Municipality name
Region	char	Region name
DimDiseaseRegister	table	Registration of infectious disease occurrence
IDRegistration	int	Key of disease registration data
IDPeriodicity	int	Foreign key from DimPeriodicity
IDLocation	int	Foreign key from DimLocation
IDFilter	int	Foreign key from DimFilter
DiseaseCategory	char	Category of the disease
Case	int	Total occurence of the disease
DimFilter	table	Filter to be applied in DimDiseaseRegister
IDFilter	int	Key of the applied filter
DiseaseName	char	Name of disease
AgeCategory	char	Category of age (0-4, 5-9, 10-14, up to over 75)
Sex	char	Sex category of disease patient
DimWeather	table	Weather observation
IDWeather	int	Key of weather observation data
IDStation	int	Foreign key from DimStation
IDTime	int	Foregin key from DimTime
Temperature	numeric	Temperature measurement (degree Celcius)

Humidity	numeric	Humidity measurement (%)
Precipitation	numeric	Precipitation measurement (mm/h)
WindDirection	numeric	Wind direction measurement (degree)
WindSpeed	numeric	Wind speed measurement (m/s)
DimAirQuality	table	Air quality observation
IDAirQuality	int	Key of air quality observation data
IDStation	int	Foreign key from DimStation
IDTime	int	Foregin key from DimTime
CarbonMonoxide	numeric	CO measurement (ug/m3)
NitrogenDioxide	numeric	NO_2 measurement (ug/m3)
Ozone	numeric	O_3 measurement (ug/m3)
PM10	numeric	PM less than 10 µm measurement (ug/m3)
PM2.5	numeric	PM less than 2.5 µm measurement (ug/m3)
SulphurDioxide	numeric	SO_2 measurement (ug/m3)
DimMarine	table	Marine observation
IDMarine	int	Key of marine observation data
IDStation	int	Foreign key from DimStation
IDTime	int	Foregin key from DimTime
WaterLevel	numeric	Water level measurement (mm)
DimStation	int	Station where the climate data recorded
IDStation	int	Key of station
IDLocation	int	Foregin key from DimLocation
StationName	char	Name of station

As seen in **Table 14**, each dimension tables have one primary key. The DimPeriodicity contains the list of periodicity which can be used for the generated report, for example, weekly, monthly or even daily. DimTime and DimLocation contains the time and location of the data. DimRegister contains the list of diseases and its number of occurrences, while DimFilter contains the list of filters to be applied to DimRegister. DimWeather, DimAirQuality, and DimMarine contain the climate data while DimStation contains the list of the Finnish observation station.

The fourth component is the data presentation. In this part, data is presented to the user for performing analysis. An example of data retrieved from the proposed data warehouse to be presented in the data presentation area is shown in **Table 15**. Furthermore, **Fig. 17** depicts the conversion of **Table 15** into a line chart.

Table 15. Example of data retrieved from proposed data warehouse

Periodicity	Location	AvgTemperature	AvgHumidity	SumCase
Yearly2019	Lappeenranta	-8.5	91	4
Yearly2019	Lappeenranta	-2	86	17
Yearly2019	Lappeenranta	1.5	81	12
Yearly2019	Lappeenranta	5.5	57	6
Yearly2019	Lappeenranta	9.8	69	4
Yearly2019	Lappeenranta	17.4	61	1
Yearly2019	Lappeenranta	15.8	70	0
Yearly2019	Lappeenranta	15.4	75	0
Yearly2019	Lappeenranta	10.2	81	0
Yearly2019	Lappeenranta	3.7	91	0

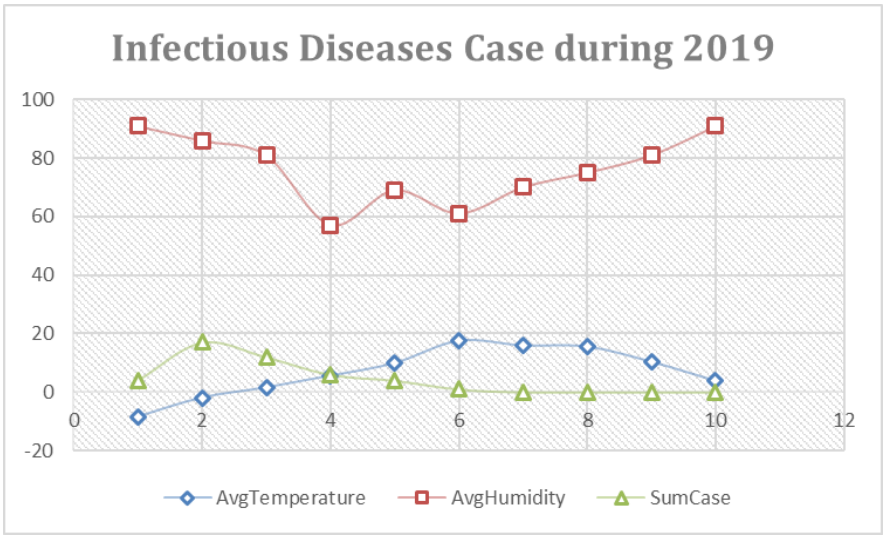

Fig. 17 Line chart converted from **Table 15**

This sample of data shown the correlation between air temperature, relative humidity, and occurrence of the infectious disease in Lappeenranta from January to October 2019 on monthly periodicity.

5 DISCUSSION AND FUTURE RESEARCH

This chapter provides the discussion part and description of related future research as continuations for this thesis work.

5.1 Discussion

Infectious disease is a term refers to all kinds of diseases caused by infectious pathogens like viruses or bacterias which distributed by person or animal. Infectious disease is considered a social problem. It often occurs as epidemics which not only harm the individual level but also on a macro scale. Many of these epidemics are linked inherently to long-term and short-term changes in climate. Various research and scientific works have proved that climate variability possesses an important role to affect the spreads of infectious diseases and their geographical distribution. Incidence of certain air-borne infectious disease like influenza and RSV commonly increase during the winter season with low temperature. Heavy rain engenders the infiltration of disease pathogens such as *Salmonella* and *Yersinia* into drinking water sources. The increase of water levels also affects some vector-borne pathogens (WHO, 2005; Chae et al., 2018).

The surveillance system is essential for prediction and controlling the growth of infectious diseases. Considering the sensitivity of climate change towards infectious disease, it is reasonable to use climate parameters, including weather, air quality and marine observation as indicators to predict the incidence of current infectious epidemics or even to forecast the growth of new disease types in a specific area. Thanks to the rapid breakthrough in technology; the availability of digital environmental and healthcare data made this means feasible also from the technical point of view. The increasing volume and accessibility of digital data, however, possess its own matter. Conforming its characteristics of volume, variety, and velocity, this big data stream is not only tremendous in its amount but also complex in structure and high at speed. These characteristics contribute as a substantial factor to the failure of the traditional system in the era of big data. It is tenuous for its poor rapidity and lack of spatial resolution. A powerful and timely

solution, therefore, is critically needed (WHO, 2005; Fang et al., 2016).

The data warehouse has become a well-established solution to resolve the passive approach to the traditional system in consolidating data from the big data stream. It is capable to integrate heterogeneous data retrieved from multiple sources into a single repository without overriding the quality and consistency requirement. Consequently, the organization may have one single source of truth where everyone can rely on the accuracy and well-managed environment to drive critical decisions. The development of a data warehouse, however, is a long and complex process which often associated with high cost. Furthermore, in some data warehouse development case, the data owner may lose control over data which inflict ownership and privacy issue. Therefore, several considerations need to be taken into account before its implementation (Sahama and Croll, 2007).

5.2 Future Research

There are several ideas for future research as a continuation of this work. It is possible to modify the existing schema to make the analysis become more generic and support more queries. Furthermore, it is also possible to add more indicators from environmental and healthcare-related data, such as radiation observation data and primary healthcare data to even more deeply delve and examine the correlation between various parameters on climate with the possibility of currently listed as well as new disease incidence. In this case, the strict security policy, of course, should be introduced for example by applying for user roles and separating data into more subset according to the user categories to ensure the privacy and confidentiality of the data owner.

6 SUMMARY

The purpose of this work is to develop a data warehouse that integrating three climate observation datasets retrieved from the Finnish Meteorological Institute: weather observation data, air quality observation data and marine observation data with the infectious disease register dataset retrieved from the Finnish Institute for Healthcare and Welfare. Only relevant parameters were selected to be included in the system, such as temperature, humidity, and precipitation. Unnecessary data were excluded during the ETL process as explained earlier. The proposed data warehouse purposely design to be used for forecasting the incidence of currently listed infectious disease as well as predicting the growth of new infectious diseases specifically in Finland. This can be done by examining the correlation between the climate parameters and disease incidence recorded on the proposed data warehouse.

The design methodology applied in this thesis work is mainly based on the theoretical framework proposed by Vaisman and Zimányi (2014). However, some modifications were also applied, considering the characteristics and limitations of the system. This design methodology consists of four closely related yet not necessarily strictly sequential phases: requirement specification, conceptual design, logical design, and physical design. The data warehouse architecture follows the bottom-up model proposed by Kimball (2013) while the requirement specification and conceptual design adopting the analysis/source-driven.

To conclude this section, it is reasonable to state that this thesis work has met its initial objective expressed in the introduction section. Furthermore, below are the set of research questions with the respective answer to support the above statement.

a. What are the roles and contributions of the data warehouse for healthcare informatics and public healthcare in general?

   Answer:
   Data warehouse possess the capacity to integrate all necessary data, be it healthcare, environmental and many other field-specific data into one centralized repository to support reporting, analysis and data-mining process supporting the healthcare

informatics and public healthcare.

b. How to integrate infectious disease register data with weather observation data, air quality observation data and marine observation data into one single repository without overriding its consistency?

Answer:
The ETL tools on staging are serving to extract clinical and climate observation data from different source systems, transform them to fulfill a set of standards and finally load the data into the data warehouse component without neglect its consistency. Furthermore, various schemas were presented to both to support as well as ensure this goal.

REFERENCES

Accenture. (2018). *Future Agenda, Open Foresight, Future of Patient Data, Insight From Multiple Expert Discussion Around the World* [PDF file]. Retrieved from: https://www.accenture.com/_acnmedia/pdf-78/accenture-health-future-of-patient-data-2018.pdf

Bansal, S., Chowell, G., Simonsen, L., Vespignani, A., & Viboud, C. (2016). Big Data for Infectious Disease Surveillance and Modeling [PDF file]. *The Journal of Infectious Diseases*, Vol. 214, Issue suppl 4, S375–S379. https://doi.org/10.1093/infdis/jiw400

Berndt, D.J., Fisher, J.W., Hevner, A.R., & Studnicki, J. (2001). Healthcare Data Warehousing and Quality Assurance [PDF file]. *Journal Computer*, Vol. 34 No. 12, 56 – 65. https://10.1109/2.970578

Chae, S., Kwon, S., & Lee, D. (2018). Predicting Infectious Disease Using Deep Learning and Big Data [PDF file]. *International Journal of Environmental Research and Public Health*, 15, 1596. https://www.mdpi.com/1660-4601/15/8/1596

Endo, N., & Eltahir, E.A.B. (2018). Prevention of Malaria Transmission around Reservoirs: An Observational and Modelling Study on the Effect of Wind Direction and Village Location [PDF file]. *Lancet Planet Health*, Volume 2 Issue 9. https://doi.org/10.1016/S2542-5196(18)30175-X

Fang, R., Pouyanvar, S., Yang, Y., Chen, S., & Iyengar, S.S. (2016). Computational Health Informatics in the Big Data Age: A Survey [PDF file]. *ACM Computing Survey* Vol. 49 No. 1 Article 12. http://dx.doi.org/10.1145/2932707

Foster, E.C., & Godbole, S. (2018). *Database Systems, A Pragmatic Approach, Second Edition* [PDF file]. Retrieved from: https://link-springer-com.ezproxy.cc.lut.fi/content/pdf/10.1007%2F978-1-4842-1191-5.pdf

George, S. (2012). *Inmon or Kimball: Which Approach is Suitable for your Data Warehouse?*. Retrieved October 25, 2019, from: https://www.computerweekly.com/tip/Inmon-or-Kimball-Which-approach-is-suitable-for-your-data-warehouse

Golfarelli, M. & Rizzi, S. (2009). Data Warehouse Design, Modern Principles and Methodology. United States: McGraw-Hill.usin

Greenwell, F., & Salentine, S. (2018). *Health Information System Strengthening: Standards and Best Practices for Data Sources, Measure Evaluation* [PDF file]. Retrieved from: https://www.measureevaluation.org/resources/publications/tr-17-225/at_download/document

Grob, M., & Hartzband, D. (2008). Health Centers and the Data Warehouse [PDF file]. *HCCN Information Bulletin* No. 14. Retrieved from: https://www.rchnfoundation.org/wp-content/uploads/2013/02/Data-Warehouse-and-Health-Centers_12_2_08.pdf

Groseclose, S.L., & Buckeridge, D.L. (2017). Public Health Surveillance Systems: Recent Advances in Their Uses and Evaluation [PDF file]. *Annual Review of Public Health Vol. 38, 57-59*. https://doi.org/10.1146/annurev-publhealth-031816-044348

Hay, I.S., George, D.B., Moyes, C.L., & Brownstein, J.S. (2013). Big Data Opportunities for Global Infectious Disease Surveillance [PDF File]. PLoS Med Vol. 10 Issue 4 e1001413. https://doi.org/10.1371/journal.pmed.1001413

Hippocrates. (1923). Hippocrates Collected Works I (W.H.S. Jones, Trans.). Cambridge, UK: Harvard University Press. (Original work published 1868). Retrieved from: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0251%3Atext%3Dintro

Inmon, W.H. (2002). *Building the Data Warehouse, Third Edition*. United States: John Wiley & Sons, Inc.

Ivančević, V., Knežević, M., Simić, M., Luković, I., & Mandić, D. (2013). Dr Warehouse - an Intelligent Software System for Epidemiological Monitoring, Prediction and Research [PDF file]. *The Fifth International Conference on Advances in Databases, Knowledge, and Data Applications*. Retrieved from: https://www.researchgate.net/publication/235352410_Dr_Warehouse_-_An_Intelligent_Software_System_for_Epidemiological_Monitoring_Prediction_and_Research

Jaakola, et al. (2017). Infectious Disease in Finland 2016 [PDF file]. Retrieved from: https://www.julkari.fi/bitstream/handle/10024/135619/URN_ISBN_978-952-302-978-1.pdf?sequence=1&isAllowed=y

Khan, S. & Hoque, A. (2015) Development of National Health Data Warehouse for Data Mining [PDF file]. *Database System Journal* Vol. VI No. 1/2015. Retrieved from: http://www.dbjournal.ro/archive/19/19_1.pdf

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit, Second Edition, The Complete Guide to Dimensional Modeling*. United States: John Wiley & Sons, Inc.

Kimball, R., & Ross, M. (2010). *The Kimball Group Reader, Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. United States: Wiley Publishing Inc.

Kimball, R. & Ross, M. (2013). *The Data Warehouse Toolkit, Third Edition, The Definite Guide to Dimensional Modeling.* United States: John Wiley & Sons, Inc.

Kristie, L.E., Hess, J.J., & Watkiss, P. (2017). *Health Risk and Costs of Climate Variability and Change, Injury Prevention and Environmental Health Third Edition* [PDF file]. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK525226/

Lane, P. (2005). *Oracle Database Data Warehousing Guide, 10g Release 2 (10.2)* [PDF file]. Retrieved from: https://docs.oracle.com/cd/B19306_01/server.102/b14223.pdf

Lane, P. (2013). *Oracle Database Data Warehousing Guide, 11g Release 2 (11.2)* [PDF file]. Retrieved from: https://docs.oracle.com/cd/E11882_01/server.112/e25554.pdf

Myers, M., Rogers, D.J., Cox, J., Flahault, A., & Hay, S.I. (2011). *Forecasting Disease Risk for Increased Epidemics Preparedness in Public Health* [PDF file]. *Adv Parasitol* Vol. 47, 2000, 309-330. https://doi.org/10.1016/S0065-308X(00)47013-2

Nichols, G.L., Andersson, Y., Lindgren, E., Devaux, I., & Semenza, J.C. (2014). European Monitoring System and Data for Assessing Environmental and Climate Impacts on Human Infectious Disease [PDF file]. *International Journal of Environmental Research and Public Health*, 11, 3894-3936. https://www.mdpi.com/1660-4601/11/4/3894

Patz, J.A., et al. (2003). *Climate Change and Infectious Diseases. Climate Change and Human Health: Risk and Responses* [PDF file]. Retrieved from https://www.who.int/globalchange/publications/climatechangechap6.pdf

Polgreen, P.M. & Polgreen, E.L. (2017). Infectious Diseases, Weather and Climate. *Clinical Infectious Diseases* Vol. 66, 815-817. https://doi.org/10.1093/cid/cix1105

Rangarajan, S. (2016). *Data Warehouse Design – Inmon versus Kimball*. Retrieved October 5, 2019, from http://tdan.com/data-warehouse-design-inmon-versus-kimball/20300

Rizzi, S. (2008). *Data Warehouse, In Encyclopedia of Computer Science and Engineering*. Italy: John Wiley and Sons, Inc.

Sahama, T.R., Croll, P.R. (2007). A Data Warehouse Architecture for Clinical Data Warehousing [PDF file]. *Conference in Research and Practice in Information Technology*, Vol. 68. Retrieved from: https://pdfs.semanticscholar.org/a6ca/8c309ba7ee7fff91e5ff5c9bff41b77b9a02.pdf

Salinas, S.O., & Lemus, A.C.L. (2017). *Data Warehouse and Big Data Integration* [PDF file]. International Journal of Computer Science and Information Technology, Vol 9, No. 2. Retrieved from: http://aircconline.com/ijcsit/V9N2/9217ijcsit01.pdf

Sanders, D., et al. (2017). *It All Starts With a Data Warehouse* [PDF file]. Retrieved from: https://www.healthcatalyst.com/wp-content/uploads/2014/02/Healthcare-Data-Warehouse.pdf

Simonsen, L., Julia R.G., Olson, D., & Vibound, C. (2016). Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems [PDF file]. *The Journals of Infectious Diseases* 2016:214 (Supplement 4). https://doi.org/10.1093/infdis/jiw376

Stanford Medicine. (2017). *Harnessing the Power of Data in Health* [PDF file]. Retrieved from: https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf

Trick, W. (2008). Building a Data Warehouse for Infection Control [PDF file]. *American Journal of Infection Control*, S76 Vol. 36 No. 3 Supplement I. https://doi.org/10.1016/j.ajic.2007.07.004

WHO. (2001). *Infectious and Infectious Disease, A Manual for Nurses and Midwives in the WHO European Union* [PDF file]. Retrieved from: http://www.euro.who.int/__data/assets/pdf_file/0013/102316/e79822.pdf

WHO. (2005). *Using Climate to Pedict Infectious Disease* [PDF file]. Retrieved from: https://apps.who.int/iris/bitstream/handle/10665/43379/9241593865.pdf?sequence=1

WHO. (2006). *Communicable Disease Surveillance and Response System, Guide to Monitoring and Evaluating, World Health Organization* [PDF file]. Retrieved from: https://www.who.int/csr/resources/publications/surveillance/WHO_CDS_EPR_LYO_2006_2.pdf retrieved 11.11.2019

WHO. (2015). *Integrated surveillance of Noncommunicable Diseases (iNCD)* [PDF file]. Retrieved from: https://ec.europa.eu/health/sites/health/files/indicators/docs/incd_en.pdf

WHO. (2018). *Air Pollution*. Retrieved November 18, 2019, from: https://www.who.int/airpollution/ambient/pollutants/en/

Wu, X., Lu., Y., Zhou, S., Chen, L., & Xu, B. (2015). Impact of Climate Change on Human Infectious Diseases: Empirical Evidence and Human Adaption [PDF file]. *Environment International*, Vol. 86, 14 – 23. https://doi.org/10.1016/j.envint.2015.09.007

Yun, S., Alsova, O., Chistyakov N., Gubarev V., Shvaykova I., & Loktev, V. (2011). Data Warehouse: Environment and Infectious Diseases [PDF file]. *Proceedings of 6th International Forum on Strategic Technology, Harbin*, 792-795. https://ieeexplore.ieee.org/document/6021140

APPENDIX 1.

Registered Infectious Diseases in Finland (Jaakola et al., 2017)

Disease Groups	**Disease Name**
Respiratory infections	Adenovirus
	Influenza
	Parainfluenza
	Rhinovirus
	Respiratory Syncytial Virus (RSV)
	Enterovirus
	Whooping cough
	Chlamydia pneumoniae
	Legionella
	Mycoplasma pneumoniae
Gastrointestinal infections	Food and water borne infection
	Clostridium difficile
	Enterohaemorrhagic Escherichia coli
	Campylobacter
	Listeria
	Salmonella
	Shigella
	Yersinia
	Norovirus
	Rotavirus
Hepatitis	Hepatitis A
	Hepatitis B
	Hepatitis C

(continues)

APPENDIX 1. (continues)

Sexually transmitted diseases	Chlamydia (chlamydia trachomatis)
	Lymphogranuloma Venereum (LVG)
	Gonorrhoea (Neisseria Gonorrhoea)
	Syphilis (Treponema pallidum)
	HIV and AIDS
Antimicrobial resistance	Methicillin-Resistant Staphylococcus Aureus (MRSA)
	Vancomycin Resistant Enterococcus (VRE)
	ESBL - Escherichia coli and Klebsiella pneumoniae
	Carbapenemase-Producing Enterobacteria (CPE)
Tuberculosis	Mycobacterium tuberculosis
Other infections	Invasive pneumococcal disease (streptococcus pneumoniae)
	Haemophilus influenzae
	Meningococcal (Neisseria meningococcal)
	Measles, Mumps, Rubella (MMR) diseases
	Varicella virus
	Borrelia (lyme disease)
	Tick Borne Encephalitis (TBE)
	Puumala virus
	Pogosta disease (sindbis virus)
	Tularemia (Francisella tularensis)
	Rabies
	Travel-related infections
	Blood and cerebrospinal fluid findings in children and adults