

Assignment 1: Problem Solving Task

A Data Warehouse for Medical data Analysis

Due date: Monday, 26 April 2021, 11:59pm

Weighting: 30% of the assessment for CAB430

Team work: You may work on this assignment individually or in pairs. In the latter case, each team member should submit an identical submission. One of the submissions will be marked. Your submission file name should be: **student1ID_student2ID.zip**. If you complete the assignment individually, your submission file name should be: **studentID.zip**.

Required to be submitted on Blackboard:

One single zip file which contains the following files:

1. A report (pdf file) containing your answers to the tasks described in Section 3 and a statement of completeness stating which tasks have been completed.
2. Your SQL Server Integration Service project (a folder containing all your project files), name your project as: **student1ID_student2ID_ETL** or **studentID_ETL**.
3. Your SQL Server Analysis Service project (a folder containing all your project files), name your project as: **student1ID_student2ID_OLAP** or **studentID_OLAP**.

1. Introduction

COVID-19 pandemic has affected everyday life worldwide. The humankind is facing unprecedented economic and social difficulties. In response to the COVID-19 pandemic, several organizations have collected some datasets and made them publicly available so that people can analyze the data in order to know more about the outbreak. Nexoid¹, a software company located in London, has created a large global medical dataset focusing on COVID-19 infection by collecting people's responses to an online survival calculator. This dataset consists of rich demographical, geographical, behavior, segmentation, health condition, and risk values. The original dataset contains about one million records. Each record contains one participant's response to the survey and two risk values calculated by the online calculator. One value is for infection risk and the other is for mortality risk.

The original dataset provides CSV or Excel files. One company wants to build a data warehouse so that they can analyze the infection risk and mortality risk more effectively for people in different regions with different health conditions or different behavior, and at different time periods. As a data analyst, you are hired by this company to build the data warehouse. You are given a subset of the original dataset as a sample dataset to build a prototype of the data warehouse.

2. Input dataset

The provided dataset consists of one CSV file, **North_America.csv**, and one Excel file, **Other_regions.xls**. The type of the Excel file is **Excel 97-2003 workbook**. **North_America.csv** provides the data collected from some countries in North America and

¹ [COVID-19 Survival \(covid19survivalcalculator.com\)](https://covid19survivalcalculator.com)

other_regions.xls provides the data collected from some countries in other regions. The two files consist of the same columns. Table 1 below lists the columns in the two files.

No	Column name	Description
1	Survey_id	Record identifier
2	Survey_date	Recorded date of the data
3	Region	Region
4	Country	Country
5	Ip_latitude	Latitude coordinate of the location
6	Ip_longitude	Longitude coordinate of the location
7	Particitant_id	Participant identifier
8	Gender	Male, Female or other
9	Age	Age quantile
10	Height	Height of the person in cm
11	Weight	Weight of the person in cm
12	Bmi	Body Mass Index
13	Blood_type	Type of the person's blood
14	Insurance	If the person has insurance or not?
15	Income	Type of income. For example, low, medium, high, or gov
16	Race	Race of the person
17	Immigrant	If the person is immigrant or not?
18	Response_id	Response identifier
19	Smoking	Information on how often the person smokes
20	Contacts_count	Number of people the person has contacted
21	House_count	Number of people living in the person's house
22	Public_transport_count	Number of people contacted by the person during public transportation
23	Working	Status of the person's work
24	covid-19_positive	0 and 1 stating the existence
25	covid19_symptoms	
26	covid19_contact	
27	asthma	
28	kidney_disease	
29	liver_disease	
30	compromised_immune	
31	heart_disease	
32	lung_disease	
33	diabetes	
34	hiv_positive	
35	hypertension	
36	other_chronic	
37	nursing_home	
38	health_worker	
39	risk_infection	Risk of the person to get infected
40	risk_mortality	Risk of the person to die

Table 1 Description of the dataset

The sample dataset is the source data for you to build the data warehouse prototype. There are some problems in the dataset, i.e., missing values, invalid value like "?", and invalid date format. You are assumed not make any changes to the source data. Instead, you are expected to correct the date format in your ETL process and not load records if they contain missing values or invalid values.

3. Tasks

(1) Source data analysis and schema design (5 marks)

You are required to design a schema for the data warehouse. The schema needs to include the necessary data about the survey. Each record in the provided source data consists of survey date, location, participant's information, participant's response to some questions, and two risk values. Your fact table should be designed for answering queries about people's infection risk and mortality risk given dimensions: date, location, participant's personal situation and responses to the survey questions. You need to consider dimension hierarchies in order to answer queries at different granularities in terms of time and location. The data warehouse cannot contain invalid values and there will be no missing data in any of the tables in the data warehouse. Moreover, no duplicate records are allowed in any table in the data warehouse.

For Task (1), you need to answer the following questions in your report:

- Study the data in the provided files and use a data profile to describe the source data.
- Design a fact table and its dimensions to model the data in your data warehouse, list the attributes in each of the dimensions.
- Choose a model (star, snowflake or fact constellation), and draw a diagram to show your proposed data warehouse schema.

(2) Destination database (3 marks)

Create a database as the destination data warehouse that implements the proposed data warehouse schema. For Task (2), you need to include in your report the following content:

- Your SQL scripts for creating your database including all the tables.
- A screenshot of the ER diagram of your database.

(3) ETL Application (14 marks)

The data warehouse will be built using Microsoft Visual Studio 2019 and SQL Server 2019. You need to extract the data from the provided source data files and build the data warehouse without missing any valid records provided in the source files.

In this task, you are required to develop an SQL Server Integration Services (SSIS) project in Visual Studio 2019 to carry out an ETL process to build the data warehouse that you have designed in the first two tasks. In your report, you need to include the following content:

- An overview of your ETL application
 - Briefly describe the purpose of your Control Flow and provide an execution screenshot of your Control Flow for your package (if you have multiple packages, you need to provide a screenshot for each of your packages). Fig. 1 shows an example execution screenshot of the control flow in the SSIS project in Week 4 practical.

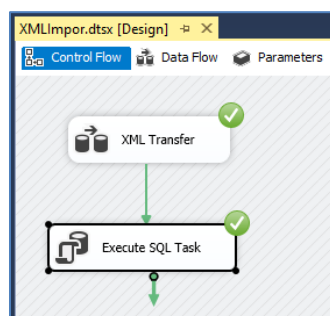


Fig. 1 An example execution screenshot of control flow in a SSIS package

- Briefly describe the purpose of each of your Data Flow tasks. You need to provide an execution screenshot for each of your data flows. Fig. 2 shows an example execution screenshot of one data flow in Week 4 SSIS project.

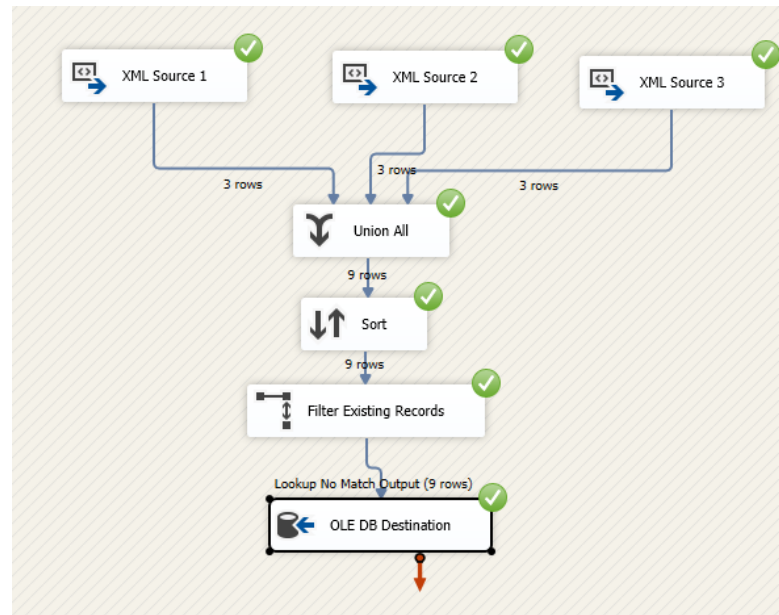


Fig. 2 An example execution screenshot of a data flow

- If you have used any SQL scripts in any of your transformations to query or modify existing database tables, please provide your SQL scripts in the report.

b) Explanation of the Transformations used in your ETL application


Describe three transformations (excluding source and destination transformations) that you have used for building the dimension tables and the fact table. Explain what each of the transformations does and why you need this transformation. Provide some screenshots to show the details if any conditions or calculations/derivations are involved.

(4) OLAP Cubes (6 marks)

The company expects that the data warehouse can provide the functionalities for them to analyze the infection risks and mortality risks in terms of different location (region or country), time (year month, or day), participants' features and their responses. At least, the company wants to be able to observe the average risks. You can choose the "AverageOfChildren" aggregation provided in the ETL tool to generate the average. Please note, this average is not an average over the records. It is an average over the days involved in the records. You also can provide data cubes that allow the company to generate maximum or minimum risk values. (The risk values in the following examples are AverageOfChildren values).

In Task (4), you are required to develop data cubes to satisfy the company's following expectations:

- Your cubes should provide three measures, i.e., the count of records (this is the default measure), infection risk, and mortality risk.
- Using your cubes, the company can observe these measures in terms of one dimension, the following figures give some examples.

Dimension		Hierarchy		Operator	Filter Expression
Date		 Date Hierarchy		Equal	{ 3, 29, 28, 29, 1, 10, 1, 8 }
<Select dimension>					

Date Year	Date Month	Date Day	Fact Survey Count	Survey Risk Infection	Survey Risk Mortality
2020	4	29	3	300	7.353
2020	4	3	2	200	0.156
2020	5	28	3	300	75.326
2020	5	29	4	400	6.484
2020	6	1	5	500	6.034
2020	6	10	1	100	0.681
2020	7	1	75	1644	80.568
2020	7	8	76	859	35.96



Dimension	Hierarchy		Operator	Filter Expression
Date		Date Hierarchy	Equal	{ 4, 7 }
<Select dimension>				
Date Year	Date Month	Fact Survey Count	Survey Risk Infection	Survey Risk Mortality
2020	4	83	745.909090909091	5.70945454545454
2020	7	1003	2013.375	95.4412500000001

Fig. 3 Results on some chosen dates or months

Dimension	Hierarchy	Operator	Filter Expression
Location	 Location Country	Equal	{ AU, US }
<Select dimension>			
Location Country	Fact Survey Count	Survey Risk Infection	Survey Risk Mortality
AU	6	43.3333333333333	0.706666666666667
US	411	380.829268292683	9.94290243902439


Dimension	Hierarchy	Operator	Filter Expression
Location	 Location Hierarchy	Equal	{ OC, EU, NA }
<Select dimension>			
Location Region	Fact Survey Count	Survey Risk Infection	Survey Risk Mortality
EU	652	331.092307692308	13.9781076923077
NA	446	403.414634146341	10.5420975609756
OC	7	35.75	0.80025

Fig. 4 Results for some chosen countries or regions


Dimension	Hierarchy	Operator	Filter Expression
Participant	 Participant Age	Equal	{ 10_20, 90_100, 60_70, 20_30 }
<Select dimension>			
Participant Age	Fact Survey Count	Survey Risk Infection	Survey Risk Mortality
10_20	51	104.045454545455	0.129272727272727
20_30	194	221.971428571429	0.395514285714286
60_70	178	139	7.26705882352941
90_100	13	134	18.6945555555556

Fig. 5 Results for some chosen participants' age groups

- c) Using your cubes, the company can observe these measures in terms of multiple dimensions, the following tables give some examples.

Dimension	Hierarchy		Operator	Filter Expression
Participant :	Participant Gender	Equal	{ female, male }	
Location	Location Region	Equal	{ OC, EU, NA }	
Date	Date Month	Equal	{ 4, 7 }	
<Select dimension>				

Date Month	Location Region	Participant Gender	Fact Survey Count	Survey Risk Infection	Survey Risk Mortality
4	EU	female	5	250	1.3755
4	EU	male	4	133.333333333333	6.08833333333333
4	NA	female	50	981	1.9674
4	NA	male	20	250	3.78075
7	EU	female	183	358	8.5335
7	EU	male	303	556.375	25.2895
7	NA	female	146	233.125	8.82675
7	NA	male	165	329.875	35.487375
7	OC	female	3	46.5	0.78
7	OC	male	4	12.5	0.41025

Fig. 6 Results in terms of time, location, and participant's feature

Dimension	Hierarchy	Operator	Filter Expression
Participant	Participant Age	Equal	{ 10_20, 40_50, 50_60 }
Response	Working	Equal	{ home, travel critical, never }
Location	Location Country	Equal	{ US, BR }
<Select dimension>			

Location Country	Participant Age	Working	Fact Survey Count	Survey Risk Infection	Survey Risk Mortality
BR	10_20	never	2	100	0.05
BR	40_50	travel critical	2	100	0.3
BR	50_60	home	1	5	0.267
BR	50_60	never	2	56.5	0.4885
BR	50_60	travel critical	2	100	0.267
US	10_20	never	1	100	0.05
US	40_50	home	14	35.6	0.4716
US	40_50	never	9	39.3333333333333	0.492
US	40_50	travel critical	9	57.1111111111111	0.283333333333333
US	50_60	home	3	36.6666666666667	0.406
US	50_60	never	16	60.5454545454545	0.778454545454546
US	50_60	travel critical	11	69.5	0.38

Fig. 7 Results in terms of location, participant's feature and behavior

You can develop one or more data cubes. The data cubes should be named using your surname followed by your data cube name (e.g., **Smith_cubeName** or **Smith_Hogan_cubeName**). In your report, you need to include the following content:

- Provide a description to each data cube which states:
 - the structure of the cube, i.e., the fact table and dimension tables,
 - what attributes are included in each dimension, and
 - attribute hierarchy if applicable.
- Provide some screenshots to support your description in a).
- Provide some screenshots of query results along with comments/explanations to show that your data cubes can satisfy the company's expectations. In the screenshots, you need to include both the Object Explorer section and the Browse section. An example screenshot of a query result is given in Fig. 8 below.

Location Country	Participant Age	Working	Fact Survey Count	Survey Risk Infection	Survey Risk Mortality
BR	10_20	never	2	100	0.05
BR	40_50	travel critical	2	100	0.3
BR	50_60	home	1	5	0.267
BR	50_60	never	2	56.5	0.4885
BR	50_60	travel critical	2	100	0.267
US	10_20	never	1	100	0.05
US	40_50	home	14	35.6	0.4716
US	40_50	never	9	39.33333333333333	0.492
US	40_50	travel critical	9	57.11111111111111	0.2833333333333333
US	50_60	home	3	36.66666666666667	0.406
US	50_60	never	16	60.54545454545454	0.7784545454545456
US	50_60	travel critical	11	69.5	0.38

Fig. 8 An example screenshot of a query result

4. Marking criteria

About your submission

- Your report should be well laid out, well-presented and easy to understand, should include student name(s) and student ID(s). There is no need for including introduction, conclusion, or references in the report. The report should just include the required contents stated in Section 3 (i.e., Tasks).
- The two projects submitted should be clearly labelled with your student ID(s). The projects will not be assessed by executing them. However, the projects will be used to verify your report content.
- Marks will be awarded for report (correctness, validity of descriptions) verified by the submitted projects.
- You will lose marks for missing or inaccurate statements of completeness.
- The values showed in your query results will not be assessed.

Note: the dataset used to generate the examples included in this assignment specification is different from the dataset provided for the assignment. Therefore, the results showed in the example figures (Fig. 3 to Fig. 8) cannot be referred as correct output for this assignment.

Marking Sheet

Student(s)

Name	Student ID	Total Marks
		/30

<u>Task 1</u>	Comments	Marks
Source data analysis		/3
The design of a fact table and its dimensions		/1
Description of the schema		/1
Sub-total mark		/5

<u>Task 2</u>	Comments	Marks
SQL statements to create a complete database including all the tables		/2
Correct ER diagram of database		/1
Sub-total mark		/3

<u>Task 3</u>	Comments	Marks
An overview of your ETL application		/8
Explain three transformations used in your ETL application.		/6
Sub-total mark		/14

<u>Task 4</u>	Comments	Marks
Description of the data cube(s)		/3
Result screenshots with brief explanations for satisfying the company's expectations.		/3
Sub-total mark		/6

<u>General</u>	Comments	Marks
Report presentation		/2
Sub-total mark		/2

--- END OF ASSIGNMENT 1 ---