**University of Westminster**

School of Computer Science and Engineering

|  |  |
| --- | --- |
| **7BUIS010W Data Warehousing and OLAP – Coursework (2023/24)** | |
| Module leader | Dr Panagiotis Chountas |
| Unit | Coursework |
| Weighting: | 50% |
| Qualifying mark | 40% |
| Description | The in-module assessment will consist of a single coursework that will assess students’ ability to utilise conceptual modelling in Data Warehouses for the needs of subject oriented analysis; it will also assess students in depth and systematic understanding of key issues, advantages and problems related to data integration and warehousing. Finally, it will assess students’ ability to conceive and implement OLAP applications, to devise effective multi-dimensional databases and to use appropriate querying languages for effective decision making. |
| Learning Outcomes Covered in this Assignment: | This assignment contributes towards the following Learning Outcomes (LOs):  LO1. critically evaluate different data warehousing and analysis approaches and their relevant benefits to business intelligence, data mining and analytics;  LO2. critically analyse the relevant merits of a data integration and data warehousing architecture;  LO3. conceptualise data modelling requirements in data warehouses for the needs of subject oriented analysis;  *.* |
| Handed Out: | 22th February 2024 |
| Due Date | 21th March 2024, Submission by 13:00 |
| Expected deliverables | Submit on Blackboard a single file containing the required documentation (either in docx or pdf format). All implemented codes should be included in your documentation together with the results/analysis. |
| Method of Submission: | Electronic submission on BB via a provided link close to the submission time. |
| Type of Feedback and Due Date: | Feedback will be provided on BB, on 12th April 2024 (appx.15 working days) |
| BCS CRITERIA MEETING IN THIS ASSIGNMENT | * 7.1.1 Critical review of literature * 7.1.2 Development of the self-directed learner * 7.1.3 Respond to opportunities for innovation * 7.1.6 Use appropriate processes * 7.1.7 Investigate and define a problem * 7.1.8 Apply principles of supporting disciplines * 8.1.1 Systematic understanding of knowledge of the domain with depth in particular areas * 8.1.2 Comprehensive understanding of essential principles and practices * 8.2.1 Produce work informed by research at the forefront * 9.1.1 Systematic understanding of knowledge at the forefront in development and implementation * of systems * 9.1.2 Comprehensive understanding of the state of the art techniques * 10.2.1 Critical awareness of current research issues, problems and/or insights |

Assessment regulations

Refer to section 4 of the “How you study” guide for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

Penalty for Late Submission

If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark, as a penalty for late submission, except for work which obtains a mark in the range 50 – 59%, in which case the mark will be capped at the pass mark (50%). If you submit your coursework more than 24 hours or more than one working day after the specified deadline you will be given a mark of zero for the work in question unless a claim of Mitigating Circumstances has been submitted and accepted as valid.

It is recognised that on occasion, illness or a personal crisis can mean that you fail to submit a piece of work on time. In such cases you must inform the Campus Office in writing on a mitigating circumstances form, giving the reason for your late or non-submission. You must provide relevant documentary evidence with the form. This information will be reported to the relevant Assessment Board that will decide whether the mark of zero shall stand. For more detailed information regarding University Assessment Regulations, please refer to the following website:<http://www.westminster.ac.uk/study/current-students/resources/academic-regulations>

Task

* 1. Task Description

# Problem specification:

Design a data mart solution for the [European Centre for Disease Prevention and Control](https://www.ecdc.europa.eu/en/covid-19/data)  (ECDC)to monitor the daily number of new reported COVID-19 cases and deaths, Intensive care unit cases and vaccination progress across the EU/EEA area.

Information regarding of **Quarantine Vaccination, Infection,** **Drug Treatment** and **Patient Recording** processes is logically described with the aid of four UML class diagram shown in Fig.1 to Fig.5.

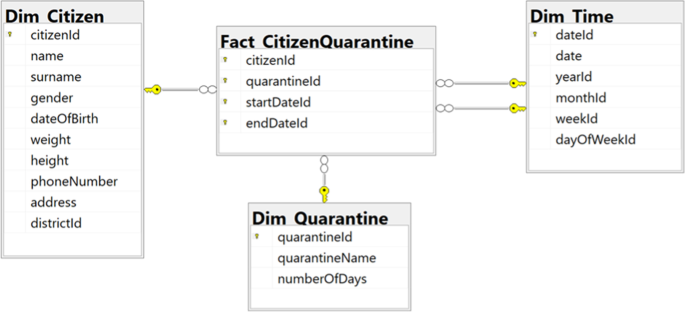


Fig. 1 Logical Design for Quarantine Process

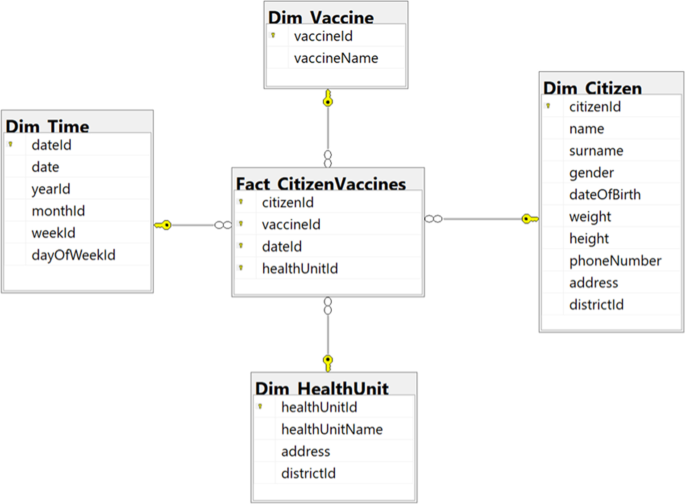


Fig. 2 Logical Design for Vaccination Process

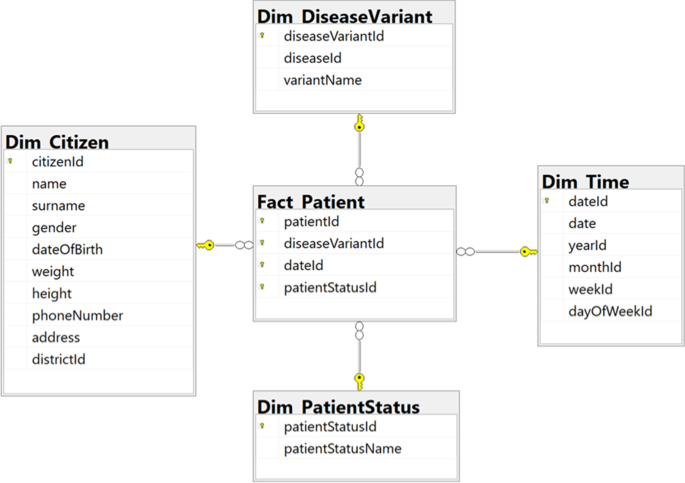


Fig. 3 Logical Design for Infection Process

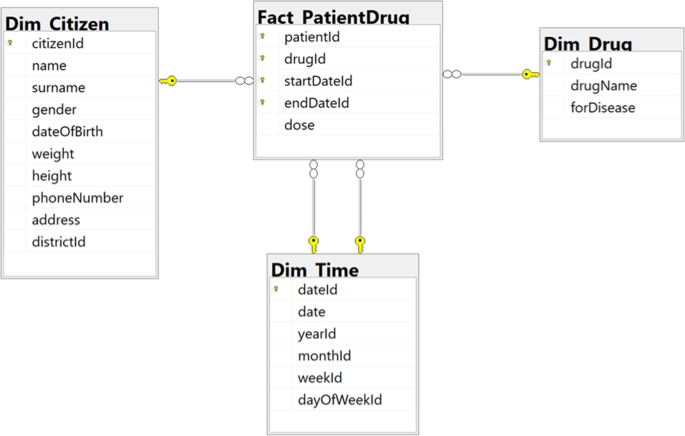


Fig. 4 Logical Design for Drug Treatment Process

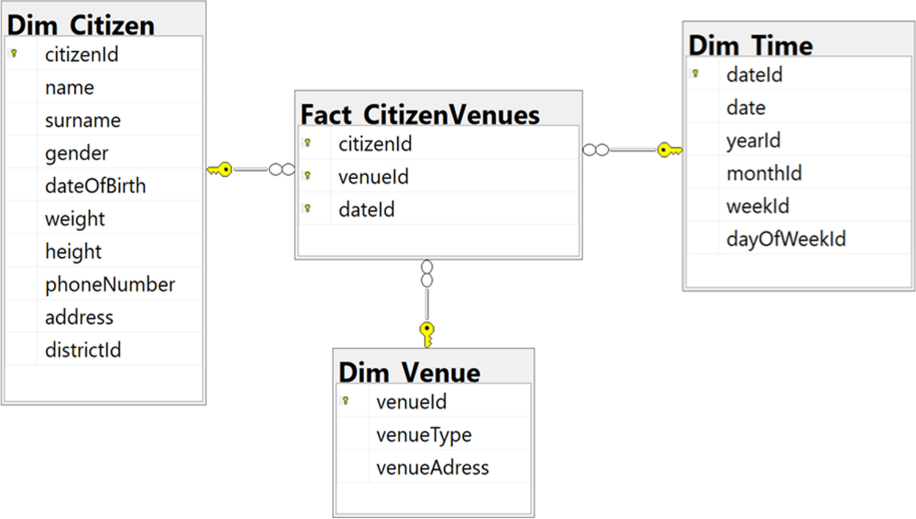


Fig. 5 Logical Design for Citizen Recording Details Process

The European health authorities would like to build a strategic data mart to answer some frequent ***queries*** the managers would like to answer are the following.

1. For each venue and month report the COVID-19 cases, vaccination rates per vaccine, quarantine and drug treatment numbers.
2. For each month report the COVID-19 cases and, quarantine and drug treatment numbers.

# Design

Considering the designed data mart and its cardinality, decide whether and which materialized views are convenient to improve response time of the frequent queries (consider all the frequent queries 1-2).

Please also consider attached information provided by [European Centre for Disease Prevention and Control](https://www.ecdc.europa.eu/en/covid-19/data)  (ECDC) regarding

* Variable\_Dictionary\_VaccineTracker
* Variable\_Dictionary\_and\_Disclaimer\_hosp\_icu\_all\_data
* Variable\_Dictionary\_and\_Disclaimer\_variant-data

Task Deliverables

1. Based on the logical schemas, in Fig.1 to Fig. 5, design an integrated dimensional fact model (DFM); in particular, the delivered DFM must promptly answer to all the *frequent queries ‘1-2’*.
   * 1. List the set of Functional Dependencies for the integrated relational schema
     2. Build the Attribute Tree based on the derived set of functional dependencies
     3. Build the Fact Schema from Attribute Tree

[16 Marks]

1. Map the DFM model to a logical model (i.e. relational). Clearly display the main fact table(s) and dimensions.

[7 Marks]

1. Implement the above logical as a working data warehouse schema, under MySQL/SQLITE, or any other suitable DBMS. Provide the DDL statements to create the proposed data-warehouse schema.

[5 Marks]

1. Considering the designed data warehouse and its cardinalities, decide whether and which materialized views are convenient to improve response time of the frequent queries (consider all the frequent queries). Explain the reasons for your choices [7 Marks]
2. Provide and implement a materialised view(s) to answer the directors *frequent queries ‘1-3’*
   1. For each venue and month report the COVID-19 cases, vaccination rates per vaccine, quarantine and drug treatment numbers. [7 Marks]
   2. For each month report the COVID-19 cases and, quarantine and drug treatment numbers. [8 Marks]

Marks for queries 5.1-5.2, will be awarded as follows: 60% for correct query formulation and 40% for appropriate display of results.

Total [50 Marks]

##### Conditions:

* + - The report must express your own conclusions and findings.
    - The overall size of your report is restricted to 20 pages, in total. Reports exceeding the 20 pages limit will be subject to a penalty of 7% out of 50, (3.5 Marks).
    - Your report should be referenced if necessary.

Marking Scheme

Due to the nature of the assessment candidates may come up with more than one equally, correct solutions. Thus marks will be allocated as follows

1. Design a conceptual data warehouse model (DFM); in particular, the designed data warehouse must promptly answer to all the *frequent queries ‘1-3’*.

[16 Marks]

* + 1. List the set of Functional Dependencies for the integrated relational schema
* *13 Primary keys corresponding to 13FD’s/ 0.5 Marks [6.5 Marks]*
  + 1. Build the Attribute Tree from the integrated relational schema

* *Construction of the Attribute Tree [3 Marks]*
* *Pruning of the Attribute Tree [2.5 Marks]*
  + 1. Build the Fact Schema (DFM) from Attribute Tree
* *Identification of Dimensions [1 Mark]*
* *Identification of Facts [1 Mark]*
* *Identification of Measures [1 Mark]*
* *Overall Correctness of delivered DFM model [1 Mark]*

1. Map the DFM model to a logical model (i.e. relational). Clearly display the main fact table(s) and dimensions. [7 Marks]

* *Mapping of Dimensions [2.5 Marks]*
* *Mapping of Facts [2.5 Marks]*
* *Identification of Measures [2 Marks]*

1. Implement the above logical as a working data warehouse schema, under SQLITE/MySQL/R/Python, or any other suitable DBMS. Provide the DDL statements to create the proposed data-warehouse schema. [5 Marks]

* *Implementation of Dimensions [1.5 Marks]*
* *Implementation of Facts [2.5 Marks]*
* *Implementation of Measures [1 Mark]*

1. Considering the designed data warehouse and its cardinalities, decide whether and which materialized views are convenient to improve response time of the frequent queries (consider all the frequent queries).Explain reasons for your choices [7 Marks]

* *Calculation of Fact table(s) size [2.5 Marks]*
* *Matrix Specification, including involved queries against Group BY and where clauses [2.5 Marks]*
* *Justified choice of Materialised views [2 Marks]*

1. Provide and implement a materialised view(s) to answer the directors *frequent queries ‘1-3*
   1. For each venue and month report the COVID-19 cases, vaccination rates per vaccine, quarantine and drug treatment numbers. [7 Marks]
   2. For each month report the COVID-19 cases and, quarantine and drug treatment numbers. [8 Marks]

* *Marks for 5.1-5.2, will be awarded as follows: 60% for correct query formulation and 40% for appropriate display of results*

Total [50 Marks]