University of Westminster

Department of Computer Science

7BUIS010W	Data Warehousing and Business Intelligence – Coursework (2020/21)
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:	18th February 2021
Due Date	18th March 2021, 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 2021 (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

1.1 Task Description

Problem specification:

Design a data mart solution for the SUN hotel chain that has over 200 hotels of different categories all over the world. On a daily basis in the OLTP system of each hotel, information on free, reserved, and unavailable rooms, booking agents and corresponding customers is stored. The hotel chain managers would like to build a data mart to analyse bookings versus checkouts and potential versus net revenue.

Information regarding bookings and payments are held in the OLTP booking system shown in Figure-1, and the e-Ticket Data Source in Figure-2. The primary keys are underlined and the foreign keys are followed by the sharp sign (#) and the name of the referenced table.

Room (RoomID, RoomTypeID#: RoomTypes, RoomBandID#: RoomBand, RoomFacilityID #: RoomFacilities, Price, Floor, AdditionalNotes)

RoomTypes (RoomTypeID, TypeDesc)

RoomBands (RoomBandID, BrandDesc)

RoomFacilities (RoomFaciliyID, FacilityDesc)

Payments (PaymentID, CustomerID#: Customer, PaymentMethodID#: PaymentMethods, PaymentAmount, PaymentComments)

PaymentMethods (PaymentMethodID, PaymentMethod)

Bookings (CustomerID#: Customer, DateBookingMade, TimebookingMade, RoomID#: Room, BookedStartDate, BookedEndDate, TotalPayementDueDate, TotalPayementDueAmount, BookingComments)

Customer (CustomerID, CustomerForenames, CustomerSurnames, CustomerDOB, CustomerHomePhone, CustomerWorkPhone, CustomerMobilePhone, CustomerEmail, CityID #: City).

County (CountyID, CountyName)

State (StateID, StateName, CountyID#: County)

City (CityID, CityName, StateID#: State)

Figure-1. Relational database schema for hotel room booking OLTP system

RoomTypes (RoomTypeID, TypeDesc)

RoomFacilities (RoomFaciliyID, FacilityDesc)

County (CountyID, CountyName)

Singer (SingerID, SingerForenames, SingerSurnames)

City (CityID, CityName, StateID#: State)

State (StateID, StateName, CountyID#: County)

Customer (CustomerID, CustomerForenames, CustomerSurnames, CustomerEmail, CityID #: City).

Room (RoomID, RoomTypeID#: RoomTypes, RoomFacilitiesID #: RoomFacilities)

Bookings (CustomerID#: Customer, RoomID#: Room, DateBookingMade, TimebookingMade, BookedStartDate, BookedEndDate, TotalPayementDueAmount)

Concert (ConcertID, ConcertName, CityID#: City, SingerID: Singer)

Buy (CustomerID#: Customer, ConcertID# : Concert, BuyDate, ConcertDate, TotalPayementB)

Figure-2. Relational database schema derived from e-Ticket Data Source

The hotel chain managers would like **integrate** the two data sources (hotel room booking OLTP system and e-Ticket Data Source)depicted in Figure-1, and Figure-2 **into a single data repository and critically analyze** the daily, monthly and yearly income. Some frequent ***queries*** the managers would like to answer are the following.

1. For each room band and month, derive the portion of rooms which are reserved, free, and unavailable.
2. For each room band, derive the portion of rooms which are reserved. Associate a rank to each county according to the portion of checkout rooms for that county in a particular year with respect to all reserved rooms for that band. The band with the highest ratio of checkout rooms in a particular year must rank first.
3. For each room band and concert, produce the cumulative income of 4-star rooms

Design

The data mart will store information from 2015 and 2019. The following cardinalities are known:

- **Room Types: ~3**
- **Hotels ~200**
- **Band: [1..5]**
- **Concerts: 2000**
- **Cities: ~500**
- **Counties: ~50**

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 1-3).

Task Deliverables

1. Merge the database schemas depicted in Figure-1 and Figure-2 into a single schema (**integrated schema**) so that can store data from both the original databases. State any assumptions you may have considered while developing the integrated schema

   [12 Marks]

2. Based on the **integrated relational schema**, design a data warehouse model (DFM); in particular, the designed data mart must promptly answer to all the *frequent queries '1-3'*.

   i. **Build the Attribute Tree from the integrated relational schema**

   ii. **Build the Fact Schema from Attribute Tree**

   [12 Marks]

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

   [6 Marks]

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

   [3 Marks]

5. 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 [4 Marks]

6. Provide and implement a materialised view(s) to answer the directors *frequent queries '1-3'*
   1. For each room band and month, derive the portion of rooms which are reserved, free, and unavailable. [4 Marks]
   2. For each room band, derive the portion of rooms which are reserved. Associate a rank to each county according to the portion of checkout rooms for that county in a particular year with respect to all reserved rooms for that county. The county with the highest ratio of checkout rooms in a particular year must rank first. [5 Marks]
   3. For each room band and concert, produce the cumulative income of 4-star rooms [4 Marks]

   Marks for queries 6.1-6.3, will be awarded as follows: 60% for correct query formulation and 40% for appropriate display of results.

Total [50 Marks]

1.2 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

Task-A 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. Merge the database schemas depicted in Figure-1 and Figure-2 into a single schema (**integrated schema**) so that can store data from both the original databases. State any assumptions you may have considered while developing the integrated schema.

[12 Marks]

- ✓ ***Identification of Tables*** ***[4 Marks]***
- ✓ ***Identification of Primary Keys*** ***[3 Marks]***
- ✓ ***Identification of Foreign Keys*** ***[5 Marks]***

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

[12 Marks]

i. **Build the Attribute Tree from the integrated relational schema**

- ✓ ***Construction of the Attribute Tree*** ***[3 Marks]***
- ✓ ***Pruning of the Attribute Tree*** ***[2 Marks]***

ii. **Build the Fact Schema (DFM) from Attribute Tree**

- ✓ ***Identification of Dimensions*** ***[2 Mark]***
- ✓ ***Identification of Facts*** ***[2 Marks]***
- ✓ ***Identification of Measures*** ***[1 Mark]***
- ✓ ***Overall Correctness of delivered DFM model*** ***[2 Marks]***

3. Map the DFM model to a logical model (i.e. relational). Clearly display the main fact table(s) and dimensions. **[6 Marks]**

- ✓ ***Mapping of Dimensions*** ***[2.5 Marks]***
- ✓ ***Mapping of Facts*** ***[2.5 Marks]***
- ✓ ***Identification of Measures*** ***[1 Mark]***

4. Implement the above logical as a working data warehouse schema, under MySQL/R/Python, or any other suitable DBMS. Provide the DDL statements to create the proposed data-warehouse schema. **[3 Marks]**

- ✓ ***Implementation of Dimensions*** ***[1 Mark]***
- ✓ ***Implementation of Facts*** ***[1 Mark]***
- ✓ ***Implementation of Measures*** ***[1 Mark]***

5. 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 **[4 Marks]**

- ✓ **Calculation of Fact table(s) size** ***[1.5 Marks]***
- ✓ **Matrix Specification, including involved queries against Group BY and where clauses** ***[1.5 Marks]***
- ✓ **Justified choice of Materialised views** ***[1 Mark]***

6. Provide and implement a materialised view(s) to answer the directors *frequent queries '1-3*

1. For each room band and month, derive the portion of rooms which are reserved, free, and unavailable. **[4 Marks]**
2. For each room band, derive the portion of rooms which are reserved. Associate a rank to each county according to the portion of checkout rooms for that county in a particular year with respect to all reserved rooms for that county. The county with the highest ratio of checkout rooms in a particular year must rank first. **[5 Marks]**
3. For each room band and concert, produce the cumulative income of 4-star rooms **[4 Marks]**

✓ **Marks for 6.1-6.3, will be awarded as follows: 60% for correct query formulation and 40% for appropriate display of results**

Total [50 Marks]