**BSc in Data Science - BSHCIFSC2**

**Advanced Databases: Project (40% of Overall)**

**Section A - Design and Implement a Data Mart - 60%**

**Section B - NoSQL Research Report - 40%**

**Due before Friday 2nd April 2021 at 23:55**

**This is a group project, maximum 2 members per group.**

There are two parts to the Assessment, **both** Section A and B must be answered by each group. Section A accounts for 60% of this assessment, Section B, 40%.

SECTION A - 60%

Using an appropriate OLAP database students are requested to implement a Data Mart by identifying an important business process and associated business requirements, designing a dimensional model that will support the requirements, implementing it using MySQL and populating it with appropriate data.

The business process should be chosen from the following list of industries and business areas. Both group members students’ numbers are to be used in this selection for the business process.

Student 1: using the last digit of your student id, please select a topic from the Industry column.

Student 2: using the last digit of your student id, please select a Business Area.

These choices must then be used to frame the business process for your data mart. Your team may implement the data mart using either a Star or Snowflake schema.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Industry** |  | **Business Area (Entity relationship)** |
| **0** | Education | **0** | Productivity (Projects – Tasks) |
| **1** | Agriculture | **1** | Customer Support (Customers – Cases) |
| **2** | Sports | **2** | Invoicing (Invoice – Items) |
| **3** | Health | **3** | Purchase Orders (PO – Items) |
| **4** | Advertising | **4** | Logistics (Parcel – Contents) |
| **5** | Leisure | **5** | Inventory (Item – Categories) |
| **6** | Retail | **6** | Payroll (Employee – Entries) |
| **7** | Finance | **7** | Recruiting (Candidate – Skills) |
| **8** | Airlines | **8** | Training (Module – Students) |
| **9** | Food | **9** | Sales (Orders – Items) |

The students should also demonstrate how the data mart meets the business requirements by the use of querying (cf. point 5 below)

In order to implement the Data Mart students are required to produce:

1. At least four important business requirements, based on the selections from the table above, for the Data Mart. The OLAP system must be able to support these requirements.
2. A dimensional model using either a Star or Snowflake schema for the Data Mart to support the business requirements. Dimensional model to clearly show the tables, column names, primary keys and foreign keys.
3. An SQL Script to create the Fact table and the Dimension tables.
4. An SQL script to load data into the Data Mart from (for example the Sakila database) and evidence of the use of manual or automatic data loading into the Data Mart. As there may be quite a lot of data that needs to be loaded into some of the tables it will be sufficient to just present a subset of the data in the document.
5. At least four queries and their outputs to demonstrate that the requirements are satisfied.

Any assumptions that are made must be stated clearly.

All this should go into **one** document and be submitted on or before **Friday, 2nd April**. Please include your student names in the filename. One report is to be submitted, with the 2 team members names to appear on the cover page.

Breakdown of marks is as follows:

Requirements 20%

Dimensional Model 20%

DDL Script 20%

DML Script 20%

Queries 15%

Student Contribution 5%

Background OLTP Database

Students may use the Sakila sample OLTP database during the investigation process for development of their data mart. The Sakila sample database was initially developed by Mike Hillyer and is intended to provide a standard schema. An ER model for the database and a description of each table can be found at:

[**https://dev.mysql.com/doc/sakila/en/sakila-structure.html**](https://dev.mysql.com/doc/sakila/en/sakila-structure.html)

The database itself can be downloaded from:

**https://dev.mysql.com/doc/index-other.html**

SECTION B - 40%

For the purpose of this report, you and your team member are expected to carry out research on a number of NoSQL Databases and write a single report. **Parts 1 and 2 should be common, part 3 specific to each student.**

The report should address the following areas:

1. Analysis and comparison of NoSQL databases with relational databases including an evaluation of the strengths and weaknesses of each as well as their future prospects.
2. Outline the motivation for the use of NoSQL databases and the main characteristics of key-value store databases, column-oriented databases, graph databases and document store databases.
3. Each team member should then describe how **two** databases works and their key features. The analysis should be approached from one of the two (odd/even) Areas of Interest.

[Parts 1 and 2 should be common, part 3 specific to each student.]

Selection Criteria for each student:

Using the **last and 2nd last** digits of your student id please select two topics from the Database Type column. Based on whether the **last** digit of your student id is odd or even please choose the appropriate Areas of Analysis and apply those 5 areas to your approach. If both last 2 numbers are the same for both students, you can still investigate the same 2 databases, but one student should choose the Odd Area of Analysis, the other student the Even Areas of Analysis.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Database Type** |  | **Areas of Analysis** |
| 0 | MongoDB | 0dd | Transactions  Recoverability  ACID/BASE Properties  Concurrency  Scalability |
| 1 | Redis |
| 2 | Cassandra |
| 3 | Neo4j |
| 4 | HBase |
| 5 | CouchDB | Even | Types of Data  Query Language  DBaaS  Cost  Security |
| 6 | InfluxDB |
| 7 | Riak |
| 8 | Memcached |
| 9 | Realm |

The report should be a MAXIMUM of 2000 words, not including Cover Page, Table of Content & Bibliography in length. Please include your student names in the filename, it should have the following characteristics:

* 12pt font / Arial / Normal Line Spacing
* Harvard Style Referencing
* Images, charts, quotations, etc. should not make up more than 25% of report
* PLAGARISM is **wrong**, please follow Library guidelines

SUBMISSION:

The final report must be submitted to Moodle on/before the deadline. Submissions received after the deadline will be subject to penalties.

MARKING SCHEME:

|  |  |
| --- | --- |
| Report Topic:   * Understanding of technical themes in topic * Relevance of chosen NoSQL databases | 10% |
| Research:   * Competency and consideration into choice of sources * Selection of appropriate sources * Quality of research * Comprehensiveness of research * Evidence of critical analysis | 35% |
| Critical Process:   * Approach towards addressing the topic * Evidence of critical reflection * Quality of argument, discussion, interpretation | 40% |
| Paper Development:   * Quality in method, organisation, interpretation and resolution of paper * Adequate and consistent referencing and citation | 15% |
|  |  |