**INSY 661: Individual Project**

**Objective**:

This individual project is designed to assess a student’s ability to understand and apply the concepts and skills learnt in the class independently as well as to go beyond what was covered in the class and learn any new concepts/skills on their own as relevant for the project requirement. It may also involve applying the programming skills (covered in other courses taken or being taken at present) to clean/manipulate/reorganize raw data available in a given format such that it may be exported/inserted in the desired database (e.g. relational or graph).

**Part 1: (80%)**

“La Ronde” is an amusement park, which is part of Six Flag’s Theme Park network, and is situated in Saint Helen’s Island in Montreal spanning 146 acres. It is the largest amusement park in Quebec and the second-largest in Canada. La Ronde has 40 rides, including ten roller coasters, of which the most popular one is named Le Monstre, considered the highest double-tracked roller coaster in the world.

La Ronde’s website has many functionalities, out of which its main two functionalities revolve around the sale of tickets for park visits and events, shopping, and relevant information for the customer like rides, park hours, etc.

The park’s business model involves the entertainment of its customers and it generates a revenue stream through the sales of daily tickets, season passes, and memberships. Apart from that, it generates additional revenue through yearly events, shopping, and food & dining.

You have joined as analyst in the company. You have been told that currently data is being stored in CSV files. You have been given a subset of the dataset (5 CSV files - ticket, ride, facility, customer, and ride\_ticket).

These files are provided to you in a zip folder - data.zip.

**You are also provided following business rules.**

The CUSTOMER file contains the customer demographic details which are necessary to uniquely identify the park visitors. A Customer can purchase one or many tickets or even has an option of not purchasing a ticket. A ticket can be purchased by one customer. Each ticket has to be purchased by one and only one customer.

The TICKET file contains the information of each ticket sold along with the type and other ticket attributes.

A ticket is necessary to access rides and facilities.

A ticket can grant access to multiple facilities (e.g., restaurants, boutiques, etc.,). A facility can be accessed by multiple ticket holders. FACILITY file contains information about each ticket used to access a facility. The same file contains detailed information about the facilities (e.e., capacity, type, location etc.)

Similarly, a ticket can be used to access multiple rides and a ride can be used by multiple ticketholders (i.e., tickets). i.e., a ticket/tickets can give you access to one or more rides. RIDE\_TICKET file contains information about each ticket used to access a ride.

The RIDES file contains the information of all the available rides and their attributes.

**Tasks**:

1. **(20%) ERD**: You have to analyse the data stored in the files and then create an ERD (identify what should be an entity, what attribute each entity should have, and how each entity should be related) to meet the given business requirements.

**NOTE:** Specify the entities, their attributes, and relationships among entities. Among the attributes of each entity, specify the primary keys. Name the relationships. Reflect on your business rules by identifying the cardinality for each relationship.

The files that you are provided may have redundant information and may suffer from update anomalies. Your ERD, if correct, would lead to normalized set of tables in the next step which will remove such duplications and eliminate update anomalies.

1. **(10%) RELATIONAL MODEL**: You should then create a relational model based on your ERD.

Use DDL to create tables in a MySQL database, as per your relational model. Link related tables together with foreign key constraints.

1. **(10%) POPULATE TABLES**: Use the data provided in the given CSV files to populate the tables that you have created in your database. NOTE: You should import all the files as tables in a MySQL database. These will act as temporary tables (youo can delete them later) that will be used to feed data (using INSERT) into tables that you have created as per your relational schema.
2. **(20%) Queries: Prepare 10 SQL queries (Complex) that will give interesting insights** (e.g., derived value, aggregate functions, etc.).

Evaluation will take into account the interestingness, usefulness, correctness, and (to a certain extent, the complexity) of the questions.

Each query should have:

1. **Clear Objective (may use concrete sample/illustration to clarify)**
2. **Assumptions (if any)**
3. **Solution i.e., SQL statement/s.**

**Example:**

Query: 1

Objective: Preferred rides in the year 2020 (preferred ride is same as popular rides which is determined based on the total number of users taking that ride).

Assumptions: A ride is considered as taken when a client has scanned his/her ticket and as a result, a ticket is associated with a ride.

Code: Select ……. ;

Output Screenshot: IMAGE

1. **(20%)** **External Data Integration and related queires**: Add external data to the existing database (for e.g., weather data) or any other data that you think might be used to provide interesting insights. You will have to do research to find our source of such dataset and figure out how to import and relate such dataset to existing dataset.

HINT: for e.g., you can use the timestamp to relate a specific facility usage to a weather condition (rain, very hot/cold temperature, etc). The matching may not be perfect. You should do your best to match such external data that has the potential to provide interesting insights. You can certainly make reasonable assumptions or state the limitations of your approach.

NOTE: Your ERD, relational model and DDL may also need to be modified.

**Prepare 5 SQL queries (Complex) that use this external data and give interesting insights**

Deliverable should include:

1. **A report** having ERD, relational model, and queries (PDF Format).
2. **A .sql file having:**
3. Properly commented SQL script for
   1. creating tables (DDLs),
   2. populating tables (i.e., importing data in supplied files and manipulating data to create desired data in the tables that are created as per the relational model).
   3. 10 queries
4. **A report** having modified ERD, modified relational model, and queries based on external data (PDF Format).
5. **A .sql file having:**
6. Properly commented SQL script for
   1. creating tables (DDLs) for external data (NOTE: please create relevant foreign key constraints),
   2. populating tables with external data.
   3. 5 queries that use the external data.

**Part 2: (20%)**

To help yourself getting ready for the job market and broadening your horizon in database-related subjects, in this part, you are asked to conduct independent research and provide a report/tutorial for your peers and yourself.

Please use one of the following topics or any other topic that you think is important to know for a data analyst/scientist. NOTE: It is a short report which should be around 5-10 pages long, so please limit the scope of the report accordingly.

|  |  |
| --- | --- |
| 1 | Query Optimization (Examples of how can we improve a query efficiency) |
| 2 | Using mongoDB : why, data modeling (compare and contrast with RDBMS), how to use it - store and query data?) |
| 3 | Using Apache Cassandra : why, data modeling (compare/contrast with RDBMS), how to use it – store/query data?) |
| 4 | Data Import/Export Using Sqoop and Flume (how to import and export data from sources, such as RDBMS) |
| 5 | Using HBase : why, data modeling (compare and contrast with RDBMS), how to use it - store and query data?) |
| 6 | AWS Redshift (what it is, why do we need it, how to use it) |
| 7 | Hive |
| 8 | Cloud Database: GCP Big Query/Azure SQL/AWS DynamoDB |

Deliverables:

1. **Content requirement:**

You should include the following perspectives in your report/tutorial:

1. The general concept of the topic/technology
   1. You need to describe the core concept and objective of the topic/technology -- Why is it needed? Alternatives? Motivation for using/learning this topic/technology?
   2. You must provide concrete examples to explain concepts (relevant and necessary coding examples and screenshots if helpful)
   3. You may describe the type of data that you can implement the selected topic/technology
2. The advantages and limitations of using the topic/technology
3. (Optional) An application of this topic/technology

NOTE: The report should be written such that it is easy to understand (even though the topic may be complex) and covers an important topic related to databse that has not been covered in class and would be helpful for a data analyst/scientist. It will be evaluated on the following parameters – lucid explanation, importance of concept, and depth in which it is covered.

1. **Format requirement:**
2. The report/tutorial is expected to be 5-10 pages long (including the cover page). Coding scripts, screenshoots and references can be put into appendix, which will not be counted towards page limit.
3. Spacing: Single-space; Font 11