**School of Computing**
**ST1501 Data Engineering CA2**
**AY2024/2025 Semester 1**

## A. Instructions and Guidelines

1. This is a group assignment with an individual component, and it accounts for **40%** of the total module grade. Each group should consist of 3 (recommended) or 4 members. Self and Peer Assessment (SPA) is used to assess individual contributions to the group component.

2. Each group is required to design and set up a data warehouse (DW) according to the business scenario described in section B using **MS SQL server**. The group's solution should demonstrate competency in designing a data warehouse and integrate data from various sources. Tasks of group component is stated in Section C, and tasks of individual component is stated in Section D. Individual component should be strictly completed by each student independently.

3. Assignment submission should be made via the ST1501 CA2 Assignment Submission link by **11:59PM, Thursday August 1st, 2024.** There are two submission tabs, 1 for group component and 1 for individual component. **Only last submission is recorded and graded.**

   **For group component:**
   Only the group leader submits a single zip file containing all deliverables of the group work via the **group submission link**. The file name should follow the convention: "**Class-GroupID.zip**" (e.g. 2B01-01.zip). Group ID will be assigned by your tutor.

   **There are 3 items in the zip file:**

   - Signed Academic Integrity Declaration Forms, one per group member.
   - A group submission template filled with the answers to Section C.
   - 4 SQL scripts:
     1. OLTP_insert.sql to insert data into the OLTP tables.
     2. DW_create.sql to create all tables in your DW.
     3. DW_insert.sql to insert data into your DW tables, including queries of necessary data from the OLTP database.
     4. DW_query.sql to answer Section C question d).

   **For individual component:**

   Each student submits a single zip file containing all deliverables of individual component via the **individual submission link**. The file name should follow the convention: "**Class-StudentNo.zip**" (e.g 2B01-P123456.zip).

   **There are 3 items in the zip file:**

- Signed Academic Integrity Declaration Form.
- An individual submission template filled with the answers to Section D.
- mongo.sql to query the data needed from the OLTP database.

4. A demo/interview will be conducted after submission deadline. During the demo session, the group is required to present its solution to explain the data warehouse design and demonstrate the process of setting up the OLTP database and the data warehouse. Each group member is required to demonstrate his/her ability to explain the data warehouse, and answer queries asked by your tutor during the presentation. Question regarding individual component may be asked.

5. Warning: Plagiarism means passing off as one's owned the ideas, works, writings, etc., which belong to another person. In accordance with this definition, you are committing plagiarism if you copy the work of another person and turning it in as your own, even if you would have the permission of that person. One shall not allow others to copy his work and such action will result in similar penalty as plagiarism. Plagiarism is a serious offence, and if you are found to have committed, aided, and/or abetted the offence of plagiarism, disciplinary action will be taken against you. If you are guilty of plagiarism, you may get 0 marks for the assignment, fail all modules in the semester, or even be liable for expulsion. As such the normal SP's academic policies on Copyright and Plagiarism applies. Please note that you are to cite all sources. You may refer to the citation guide available at: https://sp-sg.libguides.com/citation.

6. **50%** of the marks will be deducted for assignments that are received within **ONE (1)** calendar day after the submission deadline**. No marks will be given thereafter.** Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the module tutor as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

## B. The Business Scenario

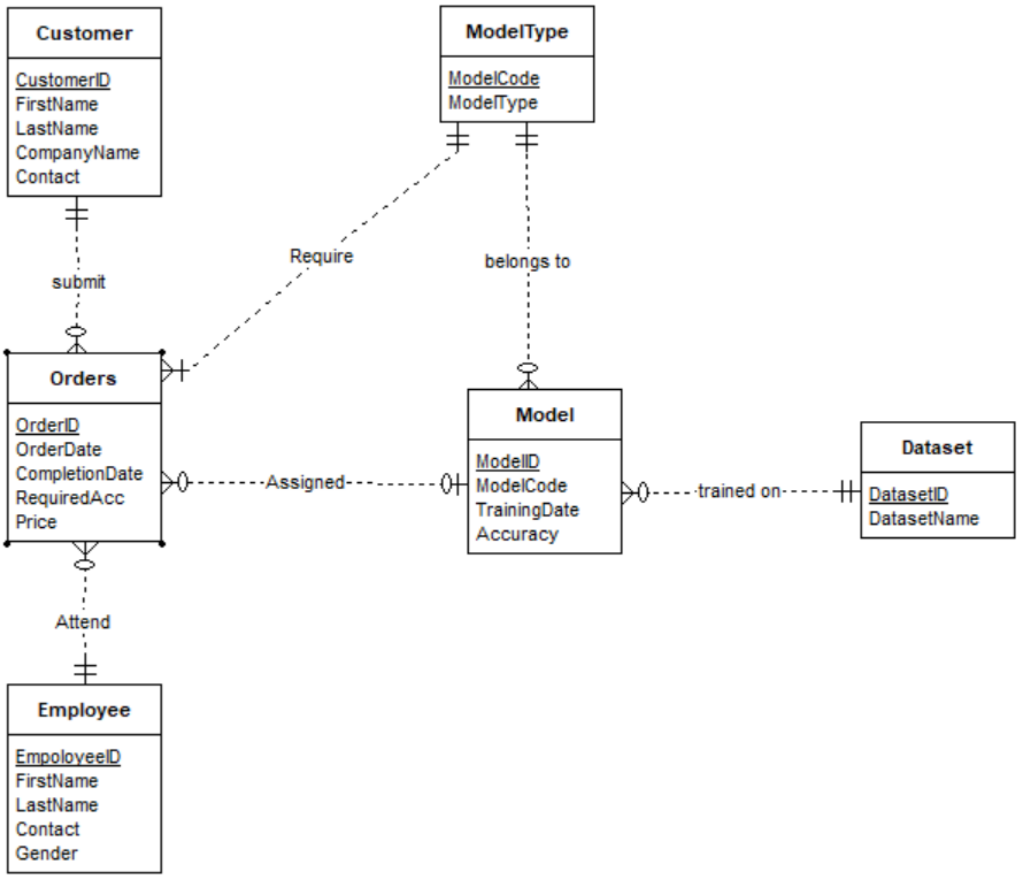
### Background

Singapore Power AI (SPAI) is a small startup company. Its main business is to provide AI solution (machine learning models) to fulfil customer orders. Currently, SPAI is using an online transaction processing (OLTP) system for business operations including tracking of model training and performance, customer orders and profits gain from completing the orders.

The owner of SPAI, Benjamin Scorp, aims to incorporate a Data Warehouse solution to support large-scale business analytics, such as seasonal trends of orders and profits. He is also interested in potential insights about customers and employees.

### OLTP Database

Below is the enhanced entity relation diagram (EERD) illustrating the current OLTP database SPAI uses. Foreign keys are not included in this EERD, and you can refer to SQL statements in appendix A on how to create these tables. **Your solution to tasks in sections C and D should be based on this OLTP strictly.**



As you can see from the above diagram, the OLTP database has 6 entities (and hence 6 tables respectively) in total. The table details are summarized below.

| Table Name | Description |
|---|---|
| Customer | Contains basic information about a customer, such as name, company name, and contact number. A unique customer ID is used as the identifier. |
| Order | This table tracks the order details when a customer submits one, with unique OrderID as the identifier.<br>Each order requests a specific type of machine learning model as the solution, with a least accuracy requirement recorded as RequiredAcc. An employee is assigned to attend the order immediately and provides a quoted price. The employee works to identify a model satisfying all the order requirements and assign it as the solution to the order. The date of assignment is logged as the CompletionDate. On the CompletionDate, the customer pays the quoted price for the order, which contributes to SPAI's profit for that day. |
| Employee | Contains information about SPAI employees, such as name, contact number, and gender. A unique employee ID is used as the identifier. |
| ModelType | This table contains unique model code and model type |
| Model | This table contains individual model that has been trained and to be assigned as a solution to orders. Each model has a unique ModelID, with detailed information such as the model code, training date and testing accuracy. |
| Dataset | This table contains information about datasets used for model training. |

## C. Group Tasks

Your group is asked to design and setup 2 databases: The OLTP database and the data warehouse implemented in MS SQL server.

**You are to make use of the data indicated in this assignment. Do not create new dataset in answering the queries.**

a) Set up the OLTP database (create the tables and insert the data) in the MS SQL Server based on the OLTP database design given in Section B, and create table SQL statements in Appendix A.

- The database should be named as SPAIXXXXYY where XXXX is replaced by your class and YY is replaced by your group ID (e.g. SPAI2A0101).
- Load data into the OLTP tables. Submit "OLTP_insert.sql" including the queries for data insertion.
- The data to be loaded to the OLTP are provided in different formats as below.
  - ➢ Customer and order information are in two separate csv files: customer.csv and order.csv.
  - ➢ Employee information is in a MS word document: employee.docx.
  - ➢ Model, dataset and model type information are in a single excel file: modeling.xlsx.
- There are a few data quality issues in the datasets provided. Identify and correct them with appropriate assumption. Describe what you have found and how you fixed them in submission template.
  Hint: Check if there are missing values/invalid values/inconsistency among data/inconsistency between data and database specification.

b) Design a star or snowflake schema for a data warehouse that will help answer various business questions based on the scenario described in Section B.

**c)** Create the data warehouse in MS SQL server according to your design in b), **after your tutor has given feedback to your design.**

- The database (DW) shall be named as SPAIDWXXXXYY where XXXX is your class and YY is your group ID (e.g. SPAIDW2A0101)
- Implement surrogate keys for all dimensions including the Time dimension and fact table(s).
- Create all tables in the DW. Submit "DW_create.sql" including all the DW table creation statements.
- Load data from OLTP tables and insert into DW tables. Submit "DW_insert.sql" including all the queries for data insertion to DW.

d) Provide meaningful queries that can be supported by your data warehouse that can draw insights regarding below questions. Submit your queries (in “DW_query.sql”), results and your discussion. Some possible insights include trend over time or other meaning dimension(s).

   1. Demonstrate insights about profits.
   2. Demonstrate insights on customer and orders.
   3. Demonstrate insights about employees.

   The queries should provide insightful findings to the owner of SPAI. You can submit up to 2 queries for each of the questions above. Explain what insights are found from query results.

## D. Individual Tasks

In the following, you shall experience how to store and manipulate data in NoSQL databases such as MongoDB. You are to show all commands and results in using MongoDB command-line only (no other languages in-built into Mongo) for each item listed below:

a) Mongo database creation
Taking reference to the **OLTP database**, create a Mongo database called SPAI with below 2 collections:

- Model – create a collection to store model details. Fields must be included are model ID, model type, training date, accuracy and total number assigned to any orders.
- Customer – create a collection to store customer information. Field must be included are customer ID, full name, company name, contact, total number of orders and total payment made.

b) Mongo Queries
Use MongoDB commands to demonstrate the following:

1. List models with total number of assignments to any orders (NoA) less than 20. List their model IDs and NoA, sorted first by decreasing NoA and then by increasing model IDs.
2. Count number of assignments and the average model accuracy per model type to 2 decimal places.
3. Enhance your query to previous question 2 to count how many model types have more than 1000 total assignments.
4. Find out customers with at least 1 order, and average payment per order is more than 450. List down full names and average payment per order of each customer.

## E. Assessment Breakdown

| Component | Task | Weightage | |
|---|---|---|---|
| **Group (60%)** | **a) Setting up OLTP database.**<br>• Create the OLTP according to the EERD in section B and create table statements Appendix A.<br>• Improve data quality before data insertion into the OLTP.<br>• Insert data into the OLTP without any error.<br>• Explain and demonstrate the correctness of OLTP setup | 10% | |
| | **b) Datawarehouse design.**<br>• The design supports business scenario in B.<br>• The chosen table names, field names and attributes are descriptive.<br>• The explanation of the design is clear and concise. | 10% | |
| | **c) Create the data warehouse.**<br>• Create the DW tables including correct primary key, surrogate key, and foreign key definition.<br>• Data inserted into DW is correct. | 10% | |
| | **d) Datawarehouse queries** | | 30% |
| | Q1 Insightful findings about profits. | 10 | |
| | Q2 Insightful findings about customers and orders. | 10 | |
| | Q3 Insightful findings about employees. | 10 | |
| **Individual (40%)** | **a) Mongo database setup** | 10% | |
| | **b) Mongo database queries** | | 20% |
| | Q1 | 5 | |
| | Q2 | 5 | |
| | Q3 | 5 | |
| | Q4 | 5 | |
| | **CA2 Assignment Demo/Interview** | 10% | |
| | **Total** | 100% | |

*** **End of Assignment Specifications** ***

## Appendix A

## OLTP table creation

```sql
create table Customer (
CustomerID varchar(10) primary key,
FirstName varchar(20) not null,
LastName varchar(20) not null,
CompanyName varchar(50) not null,
Contact varchar(10)
);

create table Employee (
EmployeeID varchar(10) primary key,
FirstName varchar(20) not null,
LastName varchar(20) not null,
Contact varchar(10),
Gender char(1) not null
);

create table Dataset(
DatasetID varchar(10) primary key,
DatasetName varchar(50) not null
);

create table ModelType(
ModelCode varchar(10) primary key,
ModelType varchar(50) not null
);

create table Model(
ModelID varchar(10) primary key,
ModelCode varchar(10) not null,
TrainingDate date not null,
Accuracy decimal(6,2) not null,
DatasetID varchar(10),
foreign key(ModelCode) references ModelType(ModelCode),
foreign key(DatasetID) references Dataset(DatasetID)
);
```

```sql
create table Orders(
OrderID varchar(10) primary key,
OrderDate Date not null,
CompletionDate Date,
RequiredAcc decimal(6,2) not null,
Price int not null,
ModelCode varchar(10) not null,
CustomerID varchar(10) not null,
EmployeeID varchar(10) not null,
ModelID varchar(10),
foreign key (EmployeeID) references Employee(EmployeeID),
foreign key (CustomerID) references Customer(CustomerID),
foreign key (ModelCode) references ModelType(ModelCode),
foreign key (ModelID) references Model(ModelID)
)
```