School of Computing

ST1501 Data Engineering CA2

Ay2020/2021 Semester 1

A. Instruction and Guideline

1. You are to work in a group of 3 (recommended) or 4 members.

2. This is a **group** assignment **with individual** component, which requires students to design and set up a data warehouse. The assignment will require students to demonstrate competency in designing data warehouse and integrate data from various sources.

3. The deadline of this assignment is on **12 August 2020 (Wed) noon.**

4. Submissions should be made via the ST1501 CA2 Assignment Submission link by the stated deadline.

5. Deliverable include:
   - An .zip file for the group report with the file-naming convention: "**ST1501-YourClass-GroupName.zip**"
   - An .zip file for the individual report with the file-naming convention: "**ST1501-YourClass-YourStudentID-YourName.zip**",

6. The group report should consist of the followings.
   a. The Database Relational Diagram of the data warehouse and the brief explanation of the design
   b. All the SQL scripts to setup the OLTP database
   c. All the SQL scripts to setup the data warehouse
   d. The four meaningful queries that can be supported by the data warehouse and its corresponding query result
   e. A presentation deck to explain the design and process of setting up the two databases.

7. The individual report should consist of the followings.
   a. All the SQL scripts to setup the data warehouse in the Databricks platform.
   b. Provide two queries as implemented in the group report in the Databrick platform and its corresponding query result
   c. Provide one addition query supported by the data warehouse setup in the Databrick platform and its corresponding query result

8. As part of the assignment requirements, the team will require to present the assignment. The presentation should explain the data warehouse design and demonstrate the process of setup the OLTP database and the data warehouse. Each team member is required to demonstrate his ability to explain the DW and the questions fielded by the tutor during the presentation.

9. Student who is absent from the presentation will be **given zero mark** for the assignment.

10. This assignment will account for **40%** of the module grade.

11. No marks will be awarded, if the work is copied or you have allowed others to copy your work.

12. 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.

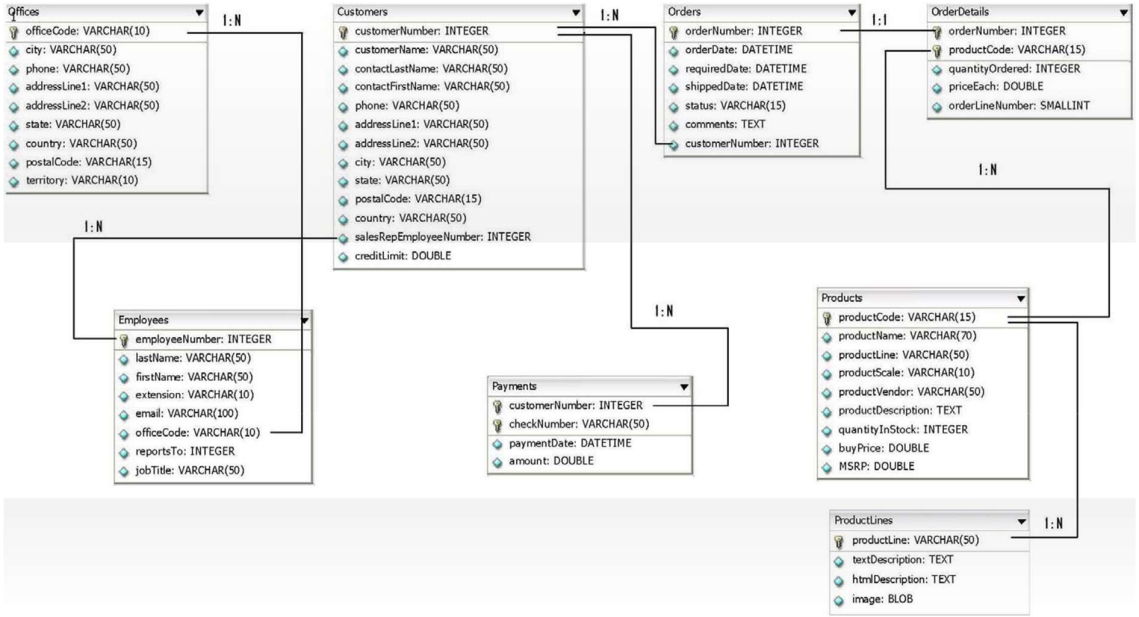
B. The Business Scenario

ClassicCars Pte Ltd is a retailer of toys specializing in scale models of classic cars. Currently, they are using an OLTP system in their day-to-day retail business operations. The data model of the database for the OLTP is shown on the next page.

The database consists of 8 tables:

1. Offices: Sales offices
2. Employees: All employees, including sales reps who work with customers
3. Orders: Orders placed by customers
4. Order Details: Line items within an order
5. Payments: Payments made by customers against their account
6. Products: The list of scale model cars
7. Product Lines: The list of product line classification
8. Customers

Recently, the CEO has decided to setup a data warehouse for analysis purpose, and he wants to explore storing all data in Databricks.


Offices
officeCode: VARCHAR(10)
city: VARCHAR(50)
phone: VARCHAR(50)
addressLine1: VARCHAR(50)
addressLine2: VARCHAR(50)
state: VARCHAR(50)
country: VARCHAR(50)
postalCode: VARCHAR(15)
territory: VARCHAR(10)
Customers
customerNumber: INTEGER
customerName: VARCHAR(50)
contactLastName: VARCHAR(50)
contactFirstName: VARCHAR(50)
phone: VARCHAR(50)
addressLine1: VARCHAR(50)
addressLine2: VARCHAR(50)
city: VARCHAR(50)
state: VARCHAR(50)
postalCode: VARCHAR(15)
country: VARCHAR(50)
salesRepEmployeeNumber: INTEGER
creditLimit: DOUBLE
Orders
orderNumber: INTEGER
orderDate: DATETIME
requiredDate: DATETIME
shippedDate: DATETIME
status: VARCHAR(15)
comments: TEXT
customerNumber: INTEGER
OrderDetails
orderNumber: INTEGER
productCode: VARCHAR(15)
quantityOrdered: INTEGER
priceEach: DOUBLE
orderLineNumber: SMALLINT
Employees
employeeNumber: INTEGER
lastName: VARCHAR(50)
firstName: VARCHAR(50)
extension: VARCHAR(10)
email: VARCHAR(100)
officeCode: VARCHAR(10)
reportsTo: INTEGER
jobTitle: VARCHAR(50)
Payments
customerNumber: INTEGER
checkNumber: VARCHAR(50)
paymentDate: DATETIME
amount: DOUBLE
Products
productCode: VARCHAR(15)
productName: VARCHAR(70)
productLine: VARCHAR(50)
productScale: VARCHAR(10)
productVendor: VARCHAR(50)
productDescription: TEXT
quantityInStock: INTEGER
buyPrice: DOUBLE
MSRP: DOUBLE
ProductLines
productLine: VARCHAR(50)
textDescription: TEXT
htmlDescription: TEXT
image: BLOB
1:N
1:N
1:1
1:N
1:N
1:N

C. The Task

As a data specialist team, your team is asked to design and setup three databases: the OLTP database, the data warehouse using MS SQL Server and the data warehouse using Databrick platform.

Group Task

(a) Design a star or snowflake schema for a data warehouse that will help answer various business questions regarding sales based on the scenario described in Section B. The team **must seek your module tutor's feedback** on the design before proceeding to setup the data warehouse.

(b) Setup the OLTP database (create the tables and insert the data) in the **Microsoft SQL Server** based on the OLTP database design given in Section B.

- The database should be named as **CarSalesXXXX** where XXXX is to be replaced by the team name.
- The SQL script to 'CREATE Table' is to be submitted
- All SQL scripts for loading data to the OLTP database is to be submitted.
- The data to be loaded to the OLTP are given as three items: Assignment_INSERT_data_partial_OLTP.sql, Products.json, Customers.csv

(c) Create the data warehouse after your tutor has feedback to your data warehouse design.

Create the database in the **Microsoft SQL Server** based on your data warehouse design.

- The database should be named as **CarSalesDWXXXX** where XXXX is to be replaced by the team name.
- Implement surrogate keys for all dimensions and fact tables.
- The SQL script to 'CREATE Table' is to be submitted

(d) Load data from the source system (the OLTP database).

- Create a temporary Time Dimension table for the data warehouse. The SQL script to generate the Time dimension table is to be submitted.
- All the SQL scripts for loading the data to the warehouse is to be submitted.

(e) Provide four meaningful query that can be supported by your data warehouse covering topics Sales, Staff/Offices, Seasons of Sales and Orders/Customer. The query should provide insightful findings to the owner. Recommended to incorporate aggregate functions, row wise functions, group by, sub-query etc.

Remember, there is no right and wrong answers, only insightful or trivial observations. That will depend on how much you want to explore and your examination of the data. Stating the obvious will not eligible for high grade. Your tutor can tell whether enough thoughts have been put into the queries presented.

- The SQL script, the corresponding business question and it query result are to be submitted.

Individual Task

(f) ~~Setup a similar database as your group~~ data warehouse in the Databricks ~~platform using your own account.~~

- The database should have all tables, **except the Time dimension*** table.
- All the SPARK SQL scripts to setup the database must be submitted.

(g) Write the SPARK SQL queries.

- ~~Provide two queries as implemented in the group report in the Databrick platform~~
- Provide one addition query can be supported by new database on the Databrick platform.
- The SPARK SQL script, the corresponding business question and it query result are to be submitted

Note: The Time dimension table is excluded for this individual task as the objective of this task is to explore the use of Databricks platform. All data warehouse should have a time dimension table.

The Assessment Criteria

Components	Weightage
Data Warehouse Design (Database Diagram) • The design supports the described business scenario. • The chosen table names, field names and attributes are descriptive. • The explanation of the design is clear and concise.	15%
OLTP Database Setup • The Create Table SQL script implements the database design, including the primary key and foreign key definition. • The Insert Records SQL script to load the given data. • The SQL Script to load the Product and Customer records • Explanation and the demonstration of OLTP setup is without error during the presentation.	15%
The database warehouse setup in MS SQL • The Create Table SQL script implements the star or snowflake design, including the primary key, surrogate key and foreign key definition. • The SQL Script to load the data into the data warehouse • Explanation and the demonstration of data warehouse is without error during the presentation.	15%
The four queries in MS SQL • Four insightful query that covers each of the category (Sales, Staff/Office, Seasons of Sales, Orders/Customer)	20%
The data warehouse setup in the Databricks (Individual) • The Create Table script to load the data to the data warehouse	15%
The three queries in the Databricks (Individual) • The two queries as implemented in the MS SQL data warehouse (5%) • One addition insightful query (5%)	10%
Question and answer during the presentation	10%

** UP to 10 marks can be deducted for poor organization of the report