Assignment 2 COS10082

NAME: Brandon Bao Quan LAI

Student ID: 102787664

Table of Contents

Part II: Data Warehouse Design 3

ERD data warehouse design: 3

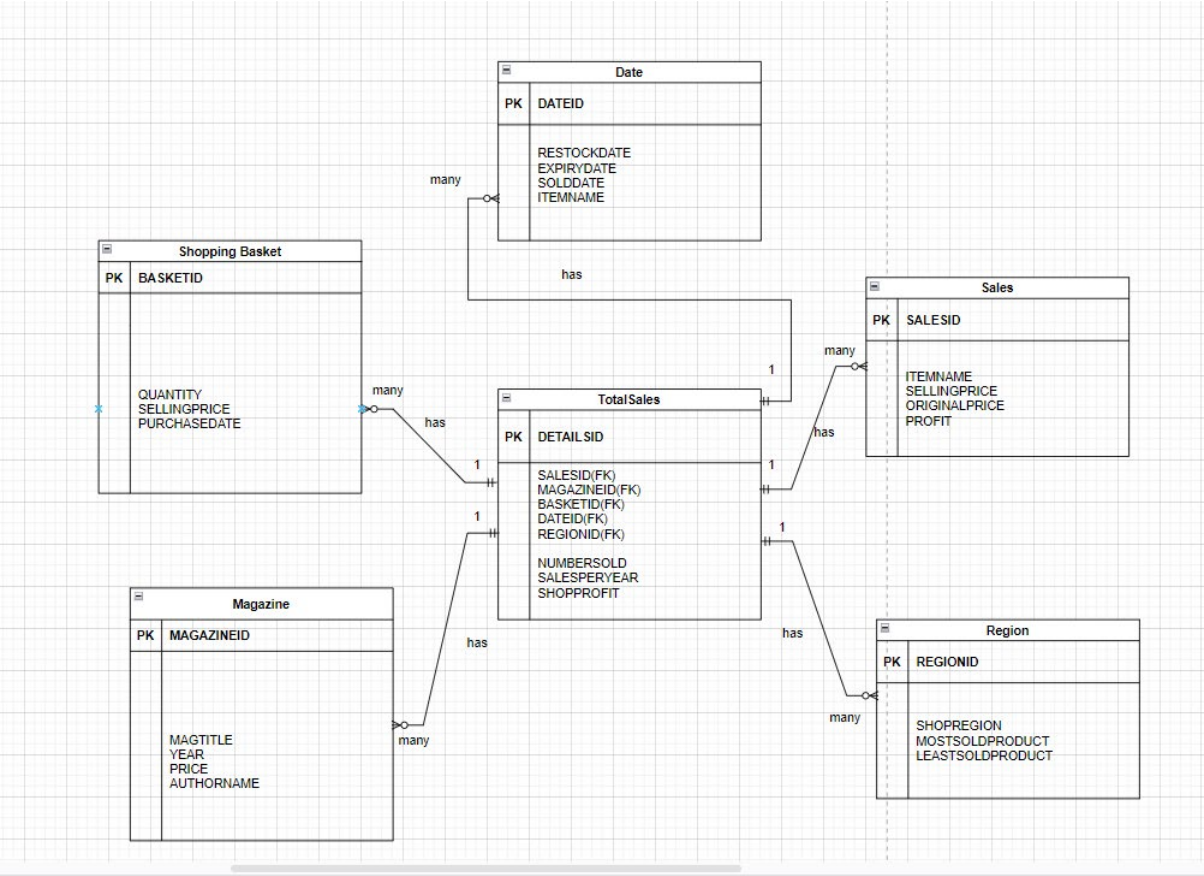
a) Provide an explanation with design justification in 500 words (±100words). Note: Marks will be deducted if the word count is not met..... 4

b) Provide a discussion on scalability and future consideration in 500 words (±100words). Note: Marks will be deducted if the word count is not met..... 6

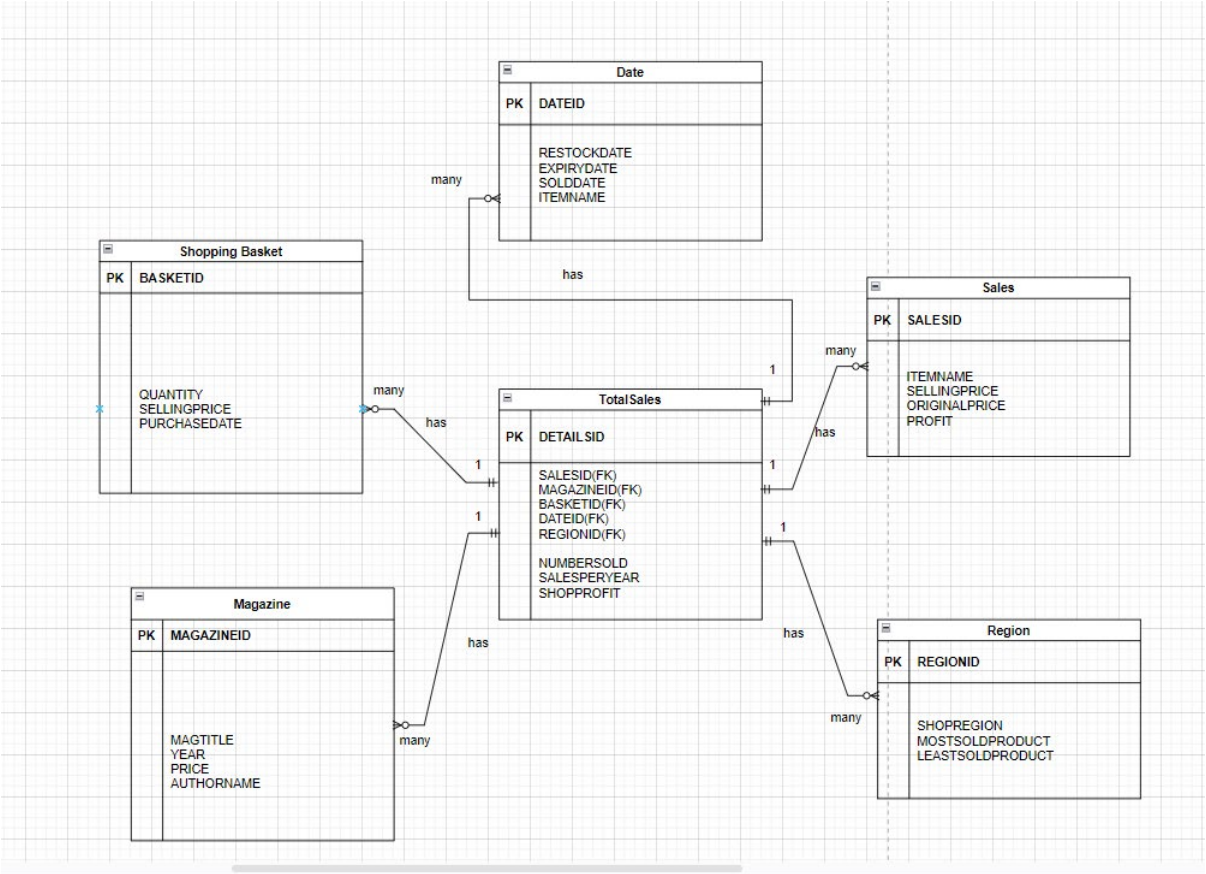
Part II: Data Warehouse Design

You are hired by Digital Dreamer to help them to bring their data from the database to a single data warehouse where the management could mainly query the data warehouse looking for data such as total sales in volume and revenue for a specific region for a particular magazine for a certain period this year compared to the same period last year. You may design the data warehouse based on the ERD in Q1

ERD data warehouse design:



a) **Provide an explanation with design justification in 500 words (±100words). Note: Marks will be deducted if the word count is not met**



The following ERD diagram was designed according to the requirements of the question. Firstly, it has the requirement met that the design has to be in a star schema design. This is shown by having a table “TotalSales” as the fact table of the star schema. It has a primary key that separates each detail by incrementing the ID of the “DETAILSID” by 1 each row it goes down the list. Each dimension table also has a primary key or surrogate key that separates each row by different number of the ID with the formula of incrementing 1 by each row. Each dimension table contains different columns of information for each unique primary key or surrogate key value, with their consecutive additional columns.

Then the primary key of each dimension table is included as the foreign key of the fact table. The details of each row of the primary key of the fact table are also done so the manager can view the total of number of products sold in a year period which is shown per year, which answers the question of showing a total volume of sales in a year compared to the same period last year. Furthermore, it can also be observed the total

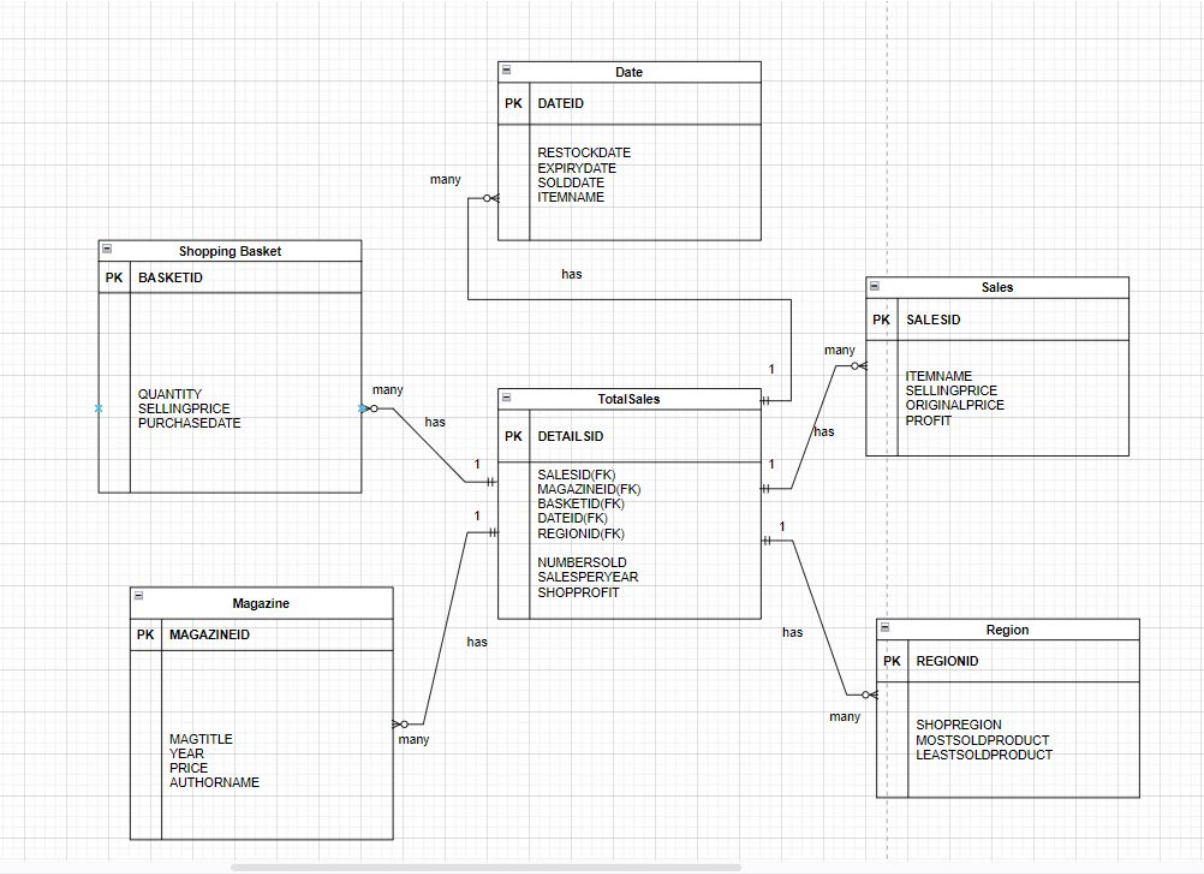
volume of sales in more than 1 year can be compared to the same period in any specific year period, allowing it to analyse the difference beyond one year. Besides the region of each total sales for a particular magazine is shown by the region ID which contains the key for shop region, most sold products, and least sold products. The region column can be conditioned with the column of sales per year column to see how each region is doing in terms of sales and number of products sold with so much more surrogate key information that is the key to other related information in each dimension table.

The number of items sold column on the fact table shows the total number of items sold, with conditions of , “SALESID” which contains, item name, selling price, original price and profit, “MAGAZINEID” which contains Magazine title, year published, price of each magazine, author’s name (taken from Question 1), “REGIONID” which contains shop region, most sold product, least sold product, “BASKETID” which contains quantity of items in it, selling price and purchase date, “DATEID” which contains the date of restock, date of expiry, sold date, and item’s name, which does not have any data viewing issue on the fact table because the result as required in the question was shown in the fact table that contains each individual meaningful row. Additionally, the mentioned columns in the fact table are conditioned with detailed non foreign key column such as sales per year and shop profit.

In conclusion, the design supports the specific query requirements of management, making it possible to analyse sales data by various dimensions (region, magazine, date). For example, a query can easily retrieve total sales volume and revenue for a specific magazine in a particular region for a defined time period, facilitating year-over-year comparisons.

(512words)

b) Provide a discussion on scalability and future consideration in 500 words (±100words). Note: Marks will be deducted if the word count is not met.



The scalability of the ERD diagram above is it could actually be implemented on a cloud server such as AWS, Azure, or Google Cloud. However, I would recommend Oracle SQL Server as it is built with a storage that is available as a cloud with SQL developer tools available to be made ready for data engineering components such as data collection, data, data storage, data processing, data pipelines and data quality and governance. Data engineering on the following database design is a backbone for the data driven enterprise, ensuring that data is accurate, reliable and accessible that provides a solid database for data analysis, machine learning, and artificial intelligence applications. The absence of data engineering on the scalability of the ERD will make the data unusable for data operations having problems such as data inconsistency, data inefficiency and unable to drive insights.

Secondly, data cleaning is recommended before doing data modelling and data analysis to prevent issues while doing the data analysis and modelling due to outliers, missing data, and inconsistent data. This could be done through the KNIME platform by using nodes, such as rule engine, missing value, and csv writer node. This could ensure consistency of data for data modelling.

Besides data modelling apps like KNIME platform is recommended for data modelling and analysis to be done on the current data of the database. Data analysis is crucial to communicate data insights from database by creating interactive visual representations, abstraction of data structures, data relationships and data rules within the system or organization. However, data modelling helps to build different models on the data to perform a lot of mathematical calculations to predict and test the accuracy of a data's predictions. This could be done through unsupervised and supervised learning models that is essential to make research or predict future markets.

After that, data analysis is also recommended to be made through libraries such as Numpy, or Pandas. Data analysis made on Pandas allow DataFrames, data manipulation, analysis and visualization. On the other hand, data analysis made on Numpy, allow arrays, matrices and high-level mathematical functions, making data manipulation and computation highly efficient. This supports better decision making, increases efficiency of work, track customer behavioural changes and identify potential risk underlying the companies' data.

Lastly, using Tableau to make data visualization is recommended as it is easy and highly efficient in creating good visualization. Data visualization allows graphs, charts, maps and many more visual representations to communicate data as well as identify potential data outliers, trends and patterns. This makes big and overwhelming data summarized and easy to understand instead of presentation of raw data. It could enhance comparison, share data efficiently, identify event relations, explore opportunities and trends. This could be done through user-centric approach, clarity and consistency, interactive and engaging visualization. Data visualization are usually the responsibility of business intelligence, reporting, analysis which could be used in sales, human resources and customer behaviour.

In conclusion, data could be stored in a cloud server, data stored could be engineered, cleaned, modelled, analysed and visualized for the scalability and future recommendations of the data warehouse ERD database.

(519 words)