

JUNGYEUN LEE

School of Computing, Science and Engineering

MSc Data Science

@00479771

# Content

1. Introduction and background: 3

1. Search Strategies, Explanation of datasets: 4
2. Data Preparations: 4-7
3. Implementation in R

* Implementation of Association rules in R: 8-11
* Implementation of Decision Tree in R: 11-15

1. Data Mining using SAS Enterprise Miner

* Association Rules Data Mining using SAS: 16-19
* Decision Tree using SAS Enterprise Miner: 19-23

1. Results Analysis and Discussion: 24-29
2. Conclusions: 29
3. References: 30
4. Index: 31-32

# Introduction and Background

A database is a structured collection of data which enable to insert, update, delete, and retrieve data easily. Traditionally databases are structured by rows and columns in individual tables within databases. Database servers allow great efficiency of performance for users who are entitled to access to the data. A database system has mainly four components: users, database application, database management system, and database.

Database Management system such as Microsoft SQL server, MySQL and Oracle is designed to perform the querying, creating and updating of data existed in relational databases to allow users for the purposes in analysis and management of the data.

Oracle and Microsoft both provide similar architecture in a concept of the relational model. However, oracle includes single instance and since database with multi-schemas whereas Microsoft offers an instance per single server which enable to create multi-databases and multi-schemas.

MS SQL server provides effectively reasonable price of ownership for both purchase and maintenance compared to other database management systems. In addition, it has been secured with the substantial degree of security along with offering a variety of scalability as organisations needed.

In this assignment, Effect Email database would be created to manage data for applications as a role of database developer. And then database administrators would connect to a user database so they execute queries which database developer have created. Users will be able to log in to SQL server with their accounts and to execute the data objects in the database with limited access depends on their permissions.

T-SQL has been adopted in the early 1970s by International Business Machines (IBM) as the new data model in the first place and it has been widely developed in the use of industrial purpose. Even though it contains several elements that allows to use in quires, it supports a restricted number of control-of-flow statements as a data retrieval language.

There are three types of T-SQL statements: data manipulation language (DML) plays a role of querying and modifying data with a use of select, insert, update, and delete;

Data definition language (DDL) deals with database objects such as tables, views, and procedures with a use of create, alter, and drop; and data control language (DCL) takes responsibilities in security permissions for users and objects in a use of grant, revoke, and deny.

From perspective of security boundaries, schemas offer the different schema level permissions and simplifies security configuration. Whilst, schemas may play a role as containers for objects as a form of tables, stored procedures, functions, and views.

# The construction of database system

As a role of a database developer consulting for a digital marketing company, effectMail schema has been generated under databases in SQL Server with total nine tables. Prior to creating those tables, entity relationship diagram has been drawn first to outline an overview as seen in the separated excel sheet.

In ERD, zero on the side of the line between two tables indicates possibility of importing non-values from value of specific table with primary key and foreign key relationship. One outlines one to one match, plus ‘n’ displays the possibility of importing multi-value matches between two tables.

Schemas are declared as fully-qualified names:

[server\_name].[database\_name].[schema\_name].[object\_name]

To create tables, each nine tables have been produced with the CREATE TABLE statement. Individual tables include a primary key with integer not null identity values. The main objective to create an index or id is security reason for a third party in case they have been exposed, also they are compulsory when a foreign key is generated. Primary key must include integer unique value supporting each row in a table. Besides, database management system automatically generates unique index once primary key is declared. Data type of identity gives an automatic value. Furthermore, not null should always have value; not allowed to leave as a blank.

Declare of ‘on primary’ in the end of statement is used in need of saving tables in specific file groups such as primary and secondary. It can be added manually through the steps of clicking the database and moving to properties.

## SQL -data types

SQL supports a wide range of data types such as exact numeric data types, approximate numeric data types, date and time data types, character strings data types, Unicode character strings data types, and binary data types.

Within exact numeric data types, several data types exist: bigint covers up to 8-byte whist tinyint takes up to 1 byte. Smallint has a scale between -32 and 32,000 whereas int has a range of values from -2 to 2 billion.

In terms of date and time, descriptions are as seen below.

| Data type | Byte | Accuracy |
| --- | --- | --- |
| Smalldatetime | 4-byte | Accuracy by a minute |
| Datetime | 8-byte | Accuracy by a milli-second |
| Datetime2 | 8-byte | Accuracy by a nano-milli-second |
| Datetimeoffset | 1 to 10-byte | Accuracy by +/- nano-milli-second |
| Date | 3-byte |  |
| time | 3 to 5-byte | Accuracy by a nano-milli-second |

In character strings data types, char outlines the fixed length whilst varchar illustrates variable length. Both enables to generate the maximum length of 8,000 characters. If there is ‘n’ in front attached to char or varchar indicates Unicode and enables to store twice over character. Varchar takes 1 byte per character (also counts space) whilst nvarchar takes 2 bytes per character (does not count space). Nvarchar is used in multi-lingual characters as well. Variable length string data type can handle to store up to 2 GB of data.

## SQL- basic query statements

* Altering tables: Tables are modified using the alter table statement. Alter table retains permissions and the data in the table.
* Dropping tables: Tables are terminated by the drop statement. All permissions and constraints can be removed apart from referenced tables such as foreign keys.
* A WHERE clause plays a role of filtering. It is used along with SELECT statement following the FROM clause. This clause is constructed from a search condition; thus, it returns as a predicate expression.

## Setting Up an effectMail Account

To set free effect account up, a user name and a secure password are created after confirmation procedure. Other information such as a valid company name, phone number and URL for the website can be filled on the later stage by calling effectMail sales team as given them ‘NULL’ value in the section of the table.

1. Create table of COMPANY with 11 columns: company index, company name, password, administration ID, first name, middle name, last name, phone number, URL for the website, email address and activation flag
2. Insert values of five companies’ details: the grill on the alley, burger and lobster, wasabi sushi, almost famous burger, and tattu.
3. Update the phone number using SET and WHERE clause

## Confirmation

When a user registers with the user email address on the website, he /she may be required to get through the email address for confirmation. As soon as the user click the button of register after filling in his /her required details, confirmation key (random numbers) is automatically created with the date-time and send to the user’s email address. When the user confirms through his /her email account, the confirmation key is matched with an email account and the confirmation active flag changes to yes (update) if the date is valid, otherwise the original confirmation key is destroyed and a new confirmation key is resent to the user.

Create table of confirmation with five attributes and build relationship between company table and confirmation table on company index through a foreign key.

CREATE PROC [dbo].[CONFIRMATION\_KEY]

@ID AS VARCHAR(100)

AS BEGIN INSERT INTO [CONFIRMATION]

(CONFIRMATION\_IDX, CONFIRMATION\_KEY, CONFIRMATION\_DATETIME)

VALUES (@ID, CONVERT(NVARCHAR(1000), CONVERT(INT, RAND() \*1000000)), GETDATE())

END

GO

C

## Multi-user account

To enables multiple users to login to the same account in giving them different permission level to access to the data, the table with relevant details in is created. Relationship between the company table and this multi-user table has been created through setting constraint of foreign key on company index. This is one of most common reflection of normalisation describes as creating separate tables for individual group of related data to organize database. Generating relationships between a new table and their predecessor using a foreign key enables to save the storage for columns which establish duplicate data in both tables with primary – foreign key relationship.

Normalisation



The process of either getting rid of duplicate data or storing adequate data in a table to minimise the space in a database. Thus, it may result in improving performance including getting a faster in speed due to smaller number of columns along with index enhancement.

In addition, as stated in introduction section, schemas provide a range of schema level to access for users by data control language mechanism in security configuration with a use of grant, revoke, and deny.

1. Build table of multi-users with 11 columns and then insert values.
2. Remove an employee who is on maternity leave and insert a replacement employee.
3. Update a temporary employee status.
4. Create another table for permission which include permission level name and description.
5. Insert values of different titles depends on different access level to the data along with descriptions.
6. Add constraint giving foreign key from permission table to multi-users table.

## Email Marketing Packages

EffectMail provides the handy calculation service for the convenience of its customers by entering the number of estimated subscribers depends on the number of subscribers. The monthly paid scheme allows clients unlimited emails to send and start from £20 per month if subscribers are lesser than 500. Next plan would be £30 per month if subscribers are up to 1,000 and then rises to £40 per month for up to 2000 subscribers.

1. Construct table of premium with variables of index, premium name, cost and description and insert values.
2. Add constraint foreign key from premium table to company table.
3. Produce stored procedure of handy calculation to obtain the price for premium customers.

CREATE PROC [dbo].[HANDY\_CALCULATOR]

@AMOUNT AS INT

AS BEGIN

IF @AMOUNT >= 2000

BEGIN SELECT 40

END

ELSE IF @AMOUNT >= 1000

BEGIN SELECT 30

END

ELSE IF @AMOUNT >= 500

BEGIN SELECT 20

END

ELSE SELECT 0

END

## Subscribers’ list

1. Establish table and insert /add values in the table, also edit /remove information.

Use of subscriber’s index instead of actual their email address. One of main reasons is in case the same subscriber has multi-accounts across other companies.

Email contacts are uploaded either manually or adding an Excel /CSV file [DB right click -> task -> import -> choosing a file source]. Be aware it should match with the structure in subscribers’ list table when those files are imported.

The case of adding a bulk of emails would be stated in the section of business intelligence tools.

Occasionally, marketing subscriber list is constructed between subscribers’ list and campaign in the real world.

A search facility in this section would be implemented in the View Section later.

## Setting up a Campaign

As including the column of user index from multi-user table, it can be checked who has created the campaign, not to mention as including the variable of company index, it can be tracked the campaign belongs to which company later.

Users can customize template by dragging and dropping elements such as images, texts, colours, and fond style. Once template is built, users can preview this on both mobile and desktop to visualize how differ on different tools. Furthermore, the template can be modified when they wish to change some components.

1. Construct template table, insert values, and edit /drop the data.
2. Build campaign table with attributes of user id, template id, campaign title, campaign text, campaign date, no errors, and sum of email numbers.

Campaign including a specific template is ready to send out to the group of selected subscribers after adding campaign title and campaign text. Prior to this, test email to the user account can be performed to see if there are any errors.

Another service of the task scheduler enables to execute tasks automatically in sending an email either at a specific time on a schedule or when a specific system occasion occurs.

As the last stage, users will be notified a pop-up message with overview of the campaign including the number of emails, campaign title, from name and from email.

CREATE PROC [dbo].[POP-UP\_MESSAGE]

@COMPANY\_IDX INT,

@USER\_IDX INT,

@CAMPAIGN\_IDX INT

AS BEGIN SELECT

c.EMAIL, ISNULL(c.FIRST\_NAME, '') + '' + ISNULL(c.MIDDLE\_NAME,'') + '' +ISNULL(c.LAST\_NAME,'') AS FULL\_NAME,

ca.EMAIL\_SUM, ca.CAMPAIGN\_TITLE FROM COMPANY c INNER JOIN [MULTI-USERS] mu ON c.COMPANY\_IDX = mu.COMPANY\_IDX

INNER JOIN CAMPAIGN ca ON mu.USER\_IDX=ca.USER\_IDX

WHERE ca.CAMPAIGN\_IDX = @CAMPAIGN\_IDX AND mu.USER\_IDX = @USER\_IDX AND c.COMPANY\_IDX = @COMPANY\_IDX

END

A searchable history of previous campaign data would be described with additional couple of stored procedure in the section of the Stored Procedure later.

## Tracking a Campaign

EffectMail has an efficient tracking system to display individual subscriber activity with a record of activity date time. The campaign status is outlined with a sum of elements such as bounces, forwards, clicks, and unsubscribes.

Particularly in this chapter of tracking a campaign, a variety of functions have been used such as system function and user defined function. System function can be performed as an expression of ISNUMERIC or ISNULL. For this task, replaces NULL with the specified replacement value. A user defined function (UDF) in SQL Server allows parameters and returns an outcome. There are several types of user defined function: for instance, scalar functions handle with a single value; and aggregate functions deal with a collection of values and returns a single value as sum of value.

* Generate table of tacking a campaign including a variety of flags such as schedule flag, bounce flag, click or forward flag with date time of its activation.

--Exercise of aggregate function

--TRACKING A CAMPAIGN: SUM OF EMAILS

SELECT

SUM (CASE BOUNCE\_FLAG WHEN 'Y' THEN 1

ELSE 0 END) AS [BOUNCE]

FROM [dbo].[TRACKING\_A\_CAMPAIGN]

WHERE CAMPAIGN\_IDX = '1'

--Use of system function

--REPLACES NULL WITH 'N' (NO) VALUE

SELECT FORWARD\_FLAG, ISNULL(FORWARD\_FLAG, 'N') AS 'FORWARD\_FLAG'

FROM [dbo].[TRACKING\_A\_CAMPAIGN]

--Use of aggregation function

--EMAIL\_SUM

SELECT COUNT(\*)[TRACKING IDX]

FROM TRACKING\_A\_CAMPAIGN

WHERE CAMPAIGN\_IDX

IN (SELECT CAMPAIGN\_IDX FROM CAMPAIGN

WHERE CAMPAIGN\_IDX = '')

# Transaction management and concurrency control

How to tackle the situation once multiple users are updating at the same time which leads to inconsistent database. To the consistency of the database, further requirements from normal programs can be carried out on transactions.

There are four key properties in transactions namely atomicity, consistency, isolation, and durability. Atomicity and consistency properties are imposed by the recovery subsystem whilst isolation property is imposed by the concurrency control subsystem.

Atomicity is referred as ‘all or nothing’ meaning that a transaction should execute the completion or does not cause any impact on the database condition. Transaction may lead to either success (‘commit’) or failure (‘abort’) that database moves on a new consistent state or it is rolled back or undone. In other words, committed transaction is unable to be aborted and an aborted transaction can be rolled back or be restarted later.

In this scenario, the property of isolation would be critical factor as one side of running transaction should not effect on other transactions in concurrency of execution. When simultaneous operations by multiple users, it may cause problems such as lost update, undone update or inadequate analysis. To prevents interference of this situation, a deadlock would be implemented in a database by locking each process requests on the resource. None of the transactions can be enforced until the other process releases.

# View

A view is a virtual table, contains rows and columns, is based on SELECT query. The view does not exist as a stored set of data values in a database. SQL join to query is common to use in the view which bring out the data from existing different tables and match the rows where related. The view is referred as the snapshot of the database.

Three sets of view are created for the convenience of enabling a search activity in the subscribers’ section.

1. Scenario one: the company of Burger and Lobster want to explore whose are resident in Manchester in their subscribers for the future promotion.

CREATE VIEW vSubscriber\_withPostcode

AS SELECT sl.company\_idx, sl.post\_code, sl.email FROM SUBSCRIBER\_LIST AS sl INNER JOIN COMPANY AS c

ON sl.COMPANY\_IDX=c.COMPANY\_IDX WHERE sl.POST\_CODE like 'M%';

GO

1. Scenario two: Almost Famous Burger company runs an event a free cocktail for their birthday subscribers.

CREATE VIEW vSubscriber\_withDOB

AS SELECT ISNULL(sl.FIRST\_NAME, '') + '' + ISNULL(sl.MIDDLE\_NAME,'') + '' +ISNULL(sl.LAST\_NAME,'') AS FULL\_NAME,

sl.company\_idx, sl.DOB, sl.email FROM SUBSCRIBER\_LIST AS sl INNER JOIN COMPANY AS c

ON sl.COMPANY\_IDX=c.COMPANY\_IDX WHERE sl.DOB = '30-10-1954';

GO

1. Scenario three: The Grill on the Alley promotes ‘Friday Blind Day Night at the Grill on the Alley’ for single customers.

CREATE VIEW vSubscriber\_withMaritalStatus

AS SELECT ISNULL(sl.FIRST\_NAME, '') + '' + ISNULL(sl.MIDDLE\_NAME,'') + '' +ISNULL(sl.LAST\_NAME,'') AS FULL\_NAME,

sl.company\_idx, sl.[MARITAL\_STATUS], sl.email FROM SUBSCRIBER\_LIST AS sl INNER JOIN COMPANY AS c

ON sl.COMPANY\_IDX=c.COMPANY\_IDX WHERE sl.[MARITAL\_STATUS] IN ('SINGLE','S');

GO

# Stored procedure

A stored procedure is a sort of code: write once and save as an object of database and then enable to execute later with writing again over and over. In other words, store a frequent query and recycle them as many as required.

In this assignment, in total four stored procedure have been generated. One was stated in confirmation section in need of producing random keys. The other one was created as a Handy Calculation to returns the value of premium cost depends on the number of subscribers. A pop-up message with overview of the campaign including the number of emails, campaign title, from name and from email was built in combination with join and where clause.

There are an additional stored procedure and statement with aggregate function outlined for a searchable history of previous campaign information below.

1. If you wish to search for campaign(s) generated by a defined user index.

CREATE PROC [dbo].[CAMPAIGN\_BY\_USER\_IDX]

@USER\_IDX INT AS SET NOCOUNT ON

SELECT \* FROM [dbo].[CAMPAIGN]

WHERE USER\_IDX = @USER\_IDX;

--AS SET NOCOUNT OFF means all the rows affected messages return from queries that executed them

--@ symbol indicates before in use of parameter

1. When you wish to display the most recent campaign from campaign table.

SELECT \* FROM [dbo].[CAMPAIGN] WHERE CAMPAIGN\_DATE=(SELECT MAX(c.CAMPAIGN\_DATE)

FROM [dbo].[CAMPAIGN] c);

GO

# SQL Server Business Intelligence tools

## SSIS

SSIS is an abbreviation for SQL Server Integration Services and is a platform for constructing efficient performance data integration solutions through extraction, transformation, and load data. SSIS package is composed with control flow engine and data flow engine.

EffectMail technical team whose support a backend technique (rather than a frontend tool) adopt a Microsoft Visual Studio tool to bring a large quantity of subscriber email list (over 20,000 contracts) into the system for premium package clients.

Steps of importing emails through a Microsoft Visual Studio

1. Tools -> add connection -> set server name and database name -> ok
2. File -> New -> project -> Integration Services Project -> change name -> ok
3. Data Flow -> Other Sources > drag adequate files such as Excel source, flat file if the data is composed with text or saved as ‘.csv’, or OLE DB source if the data is generated in SQL -> select and drag OLE DB Destination
4. Double click on flat file source box -> add the file with emails’ in -> click ‘advanced’ and modify data type if applicable
5. Double click on OLE DB source -> connect to the database and select specific table name -> check on ‘Mappings’
6. Connect a line between two boxes and check on the section of control flow letting them to loop through the data before run

## SSRS

SQL Server Reporting Service offers data visualization forms using bars, charts, spark-lines and maps

Microsoft SQL server 2014 Reporting Services Configuration Manager

Email Settings -> set sender address and SMTP Server

# Advice on database security

# References

<https://technet.microsoft.com/en-us/library/ms191178(v=sql.105).aspx> normalization

<https://msdn.microsoft.com/en-us/library/aa291817(v=vs.71).aspx> normalization

<https://docs.microsoft.com/en-us/sql/relational-databases/user-defined-functions/create-user-defined-functions-database-engine>

<http://www.serversmtp.com/en/smtp-configuration> confirmation

<https://msdn.microsoft.com/en-us/library/windows/desktop/aa383614(v=vs.85).aspx> scheduler

<https://technet.microsoft.com/en-us/library/ms169917(v=sql.110).aspx> ssis