CHAPTER 3
RESEARCH METHODOLOGY

3.1 Literature Study

The first step to do this research is doing literature study. This project start by searching for scientific journals and articles published in the internet. By searching the scientific journals and articles, the material sought are state-of-the-art related to deep neural network. After this step is done, there is a clear foundation about the way for implementation.

3.2 Simulation Database

To create simulation database, this study need to understand how the data are stored in the database and the general pattern of traffic in the production database. After the general understanding about the production database achieved, the design of the simulation program can be made.

After designing the flow of the program, the next step is to create it. Oracle Database is used as it's customizable to limit the size of objects in the database, especially limit the size of tablespace by manually adding datafile. This production database is used as base dataset for testing the neural network.

The data will be acquired from the simulation database by creating one main program. The tasks of the main program is to run the simulation database (adding data to tablespace) and fetching the data each iteration. Each iteration add x amount of data in table inside the tablespace and fetching tablespace data for the current iteration. This iteration means each time the main program runs each day. The main program iterating 5 times in a day (9 a.m., 12 noon, 3 p.m., 6 p.m. and 9 p.m.). This program runs until the tablespace is full (counted as a cycle). In the

end, *n* (total cycle per tablespace) worth of datasets created to test the neural network. The fetched data are: date, time, size (maximum size and current size of the tablespace), and date when the cycle start.

3.3 Analyzing and Designing the System

The purpose of this step is to analyze current system to forecast full storage of database. After analyzing the current system is done, new better system is designed and after it's done the new system can be created.

3.4 Implementing the system

The full system design will be discussed in chapter 4. This section discuss the summary of the system.

This system use python as the programming language. Oracle Database used as corpus database to be tracked and saving the dataset before being processed later(when coverting to CSV). For the neural network, framework keras with tensorflow as a backend is used because the system need to save and load the neural network model, using the available optimization, and quick development process.

This new system consists of 3 main parts:

1. Mine Dataset Process

   This is the main process of the system. First before running the entire system, the database administrator need to run the settings. This settings is responsible to set the threshold size when the database considered full, alert (using email) and connection to database. After that, database administrator only need to set cron job to run the main program and let it run automatically.

2. Training process

   This proccess get invoked if the mine dataset process detect that the tablespace size reach the threshold. This process train the neural network with newest data each time it got invoked. The neural network itself use keras as the framework. All layer in the network use Leaky ReLU as activation function because it deal with regression problem (output value in days). Mean Square Error used as cost function. For optimizer, it use adam since it prove to makes faster progress [17]. After it's done training, the model is saved for use later (for future training or predicting).

3. Predicting process

   Administrator can run this process to predict days left until the tablespace is full. For predicting, having saved model inside the model storage is mandatory. This process load the saved model and run it against the data of the current date the program get launched.

3.5 Testing and Tweaking the Neural Network

In this stage, the neural network get tested and various parts get adjusted. For testing, graph is used after training and predicting the data fetched from simulation database to see it's performance. This study use correct prediction and one-off day in prediction as accurate (if the prediction is 36.8 which get rounded to 37 and the real days until full is 36, it's treated as correct prediction/accurate), other than that, it's treated as inaccurate prediction.

The parts in the network that get tweaked are:

a) Inputs

   Processed data that's being fed as input to neural network changed accordingly to acquire the optimal result.

b) Neurons and Layers

Changing the neurons affect the accuracy of the neural network. Neural network neurons and layers must be tweaked so it doesn't overfit and predict correctly with the new unseen data.

c) Activation Function

Finding the most optimal alpha in leaky ReLU needs trial-error observation. Optimal value for alpha is different for each neural network and the problem it's trying to solve. Accuracy get affected by activation function.

d) Epoch and Batch

Changing epoch and batch greatly affect the accuracy of the neural network. Batch size means number of sample/data processed before updating the model. Epoch means number of loop passes through the entire dataset. For example if batch size is 10, the program go through 10 sample in dataset, calculating the error, then update the model. It means it's going for the entire epoch, and update it every 10 sample. Deciding batch size depends on how much sample in dataset and how often it needs to update. Deciding number of epoch depends on activation function it use. In a test, sigmoid needs more epoch to reach small improvement in the prediction, but it is not the case for tanh and relu [19].