CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

1. Oracle Database Storage

   The Oracle Database store data in form of physical file called datafile. A datafile belong to a tablespace, and a tablespace hold one or more datafile. Tablespace can hold more than one table in different schema (user). To add space in tablespace, database administrator can either add new datafile(s) or resize the current available datafile(s) inside the tablespace. Oracle also give some options to add these space, whether to use available auto extend function or add it manually.

   Even though the auto extend method can lift some database administrator's work, it still give some problem. Few of them are:

   Oracle leave no free remaining space for the data to grow.

   • Application or user error can fill up the whole database.

   • Lack of file system control.

   • The process of extending datafile(s) causes performance issue

   For that reasons, database administrator usually manually add datafile and monitor the database.

2. Current Oracle Database Storage Administrating Process

   The current oracle database storage administrating process which use self-developed alert script goes as follows (see Illustration 4.1):

   Admin install database and prepare everything in production database, including tablespaces and data allocations. This database takes all the

records of transactions that the user made. Then admin initialize the script to alert if space usage is above certain threshold.

If this script detect that space usage is reaching set threshold, it alerts the administrator. Then the administrator need to take further action by adding datafile inside tablespace or leave them be.

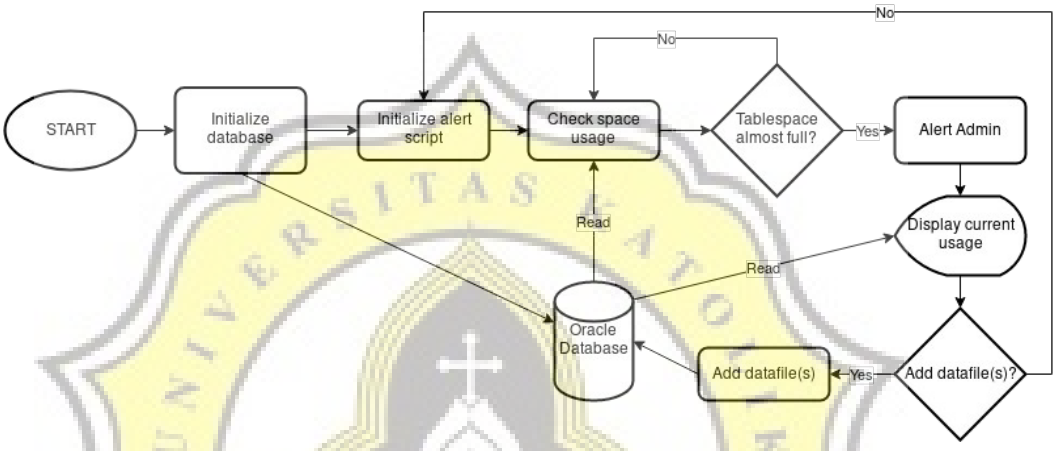

Illustration 4.1: Current oracle database storage administration process

a) Problem with current system

With manual monitor and forecasting of storage system, it requires administrator to have enough experience to predict when the database storage will be full. For new administrator who haven't adapt to the current workplace or database usage pattern will have hard time to predict when the database storage will be full.

Another problem is that the administrator need to constantly monitor the space usage to make sure it doesn't go over the safe range. If the administrator doesn't need to constantly monitor the space usage (using application to alarm usage), it will be problematic if the administrator unable to control the database (day off, sick, etc) and doesn't prepare counter-measure beforehand because they can't predict the space usage.

The problem with alerting system is it take the maximum size of the tablespaces as it's parameter for alerting the database administrator. For example, if the database add 10 GB for each datafile, and the database

administrator set 90% of maximum size in tablespace, which that would be 9 GB for alert (1 GB of free space remaining). But if administrator keep adding 10 GB until the maximum size reach 1000 GB, the alerting system alerts the administrator if the size reach 90% of that size which is 900 GB (100 GB of free space remaining). This is a lot of free space remaining, but the system still give a warning to the admin.

b) Summary of the current system

The storage administrating process in Oracle Database have two option to extend the datafile, automatically or manually. The manual used in the system to make the database administrator have full control over the storage and prevent error which will fill up the whole database where auto extend does.

This system need to have data administrators who have enough knowledge and experience to forecast the full storage of the oracle databases. New administrator won't know when the database will be full, and can't take proper counter-measure beforehand.

This system use alerting system which take percentage of current maximum system as parameter. This system will be a problem as the tablespace size continues to grow, because of the disparity between the threshold on the start of tablespace and when it grows.

c) New System

New system add a forecasting system of database storage, which in turn try to solve the problem in the current system. It acts to predict when the database will be full(in days). This will give the database administrator some time to prepare few days prior to make counter-measure (for example adding datafile(s), resizing datafile(s), etc). Another benefit this will give is the database administrator has less time to constantly monitor the database storage and focus on other task.

This system implements machine learning algorithm(neural network) to make a prediction. This technique is used because prediction depends on the aspects of experience to train the model. Another thing that affecting this technique is relatively constant input of data that production database takes each range of time. With the relatively constant flow of data that database takes, it creates a pattern which is useful to forecast full storage. Then this system should predict the time in days when the database will be full.

4.2 Design

1) Database Schema

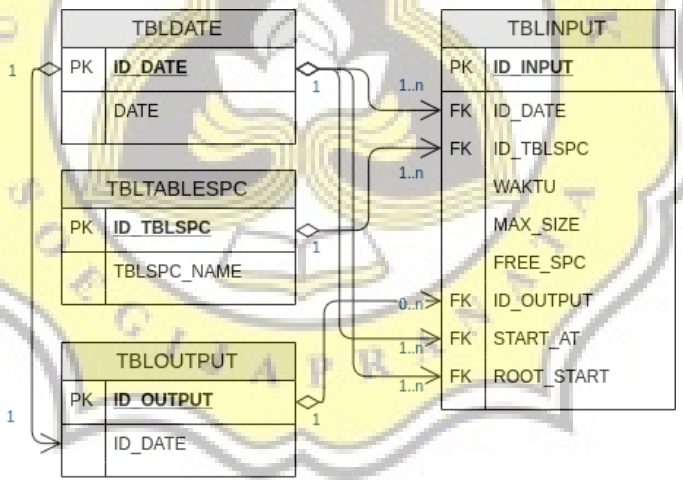

Illustration 4.2: System database schema

Database schema for the system consists of 5 tables linked together. This schema ensure all raw data are stored and can be logically easier to store. Dataset can be easily modified to processed data with this raw data.

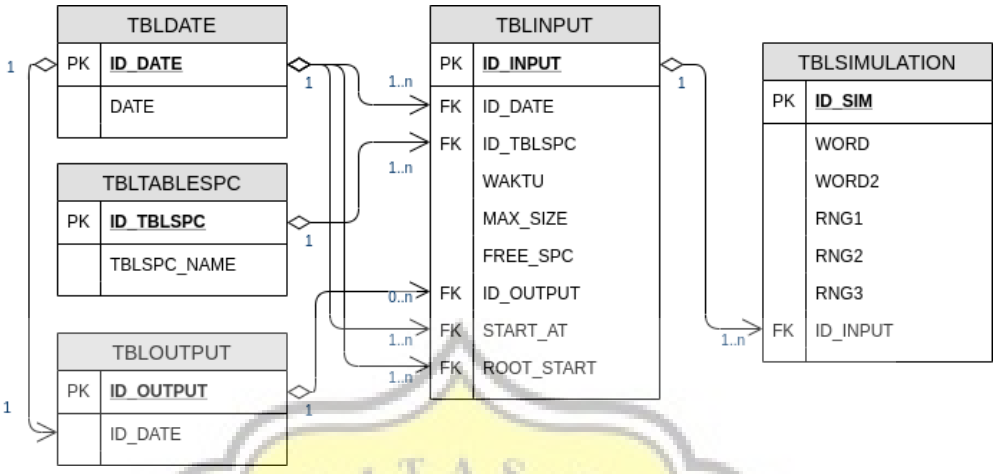

Illustration 4.3: Testing database schema

The difference between the system schema and testing schema is the table for simulation. This table simulation is used for the tracked tablespace storage.

2) Core Processes Summary

The system have 3 main/core process: mine data set process, training process and predicting process.

a) Mine dataset process

This process is tasked for fetching and processing required data to train the neural network. This parameter takes some data from the oracle database. Before running this process and other process, settings program must be done.

b) Training process

This process is tasked for training the data acquired from mine dataset process. Runs only after it got invoked by mine dataset process.

c) Predicting process

This process is tasked for predicting current tablespace size using model trained by training process.

3) Processes Exposition

a) Mine dataset process

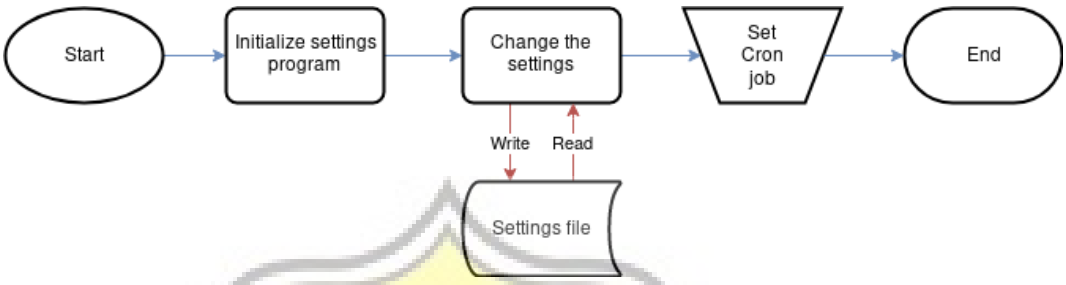

Illustration 4.4: Change the settings

Before we run the entire process, we need to set the settings first. Illustration 4.4 shows the flow of the setting preparation. We need to run this once before running the other process first especially the mining process.

There will be a program to set things up and prompt the user (database administrator) for some settings. This program changes the parameters in the settings file accordingly.

After the user made the changes in the settings, the user need to set the cron job to run mine dataset program at certain time of the day. When this entire flow is done, the program does all the work automatically.

This mine dataset process tasks is to manage dataset that will be used as training for neural network.

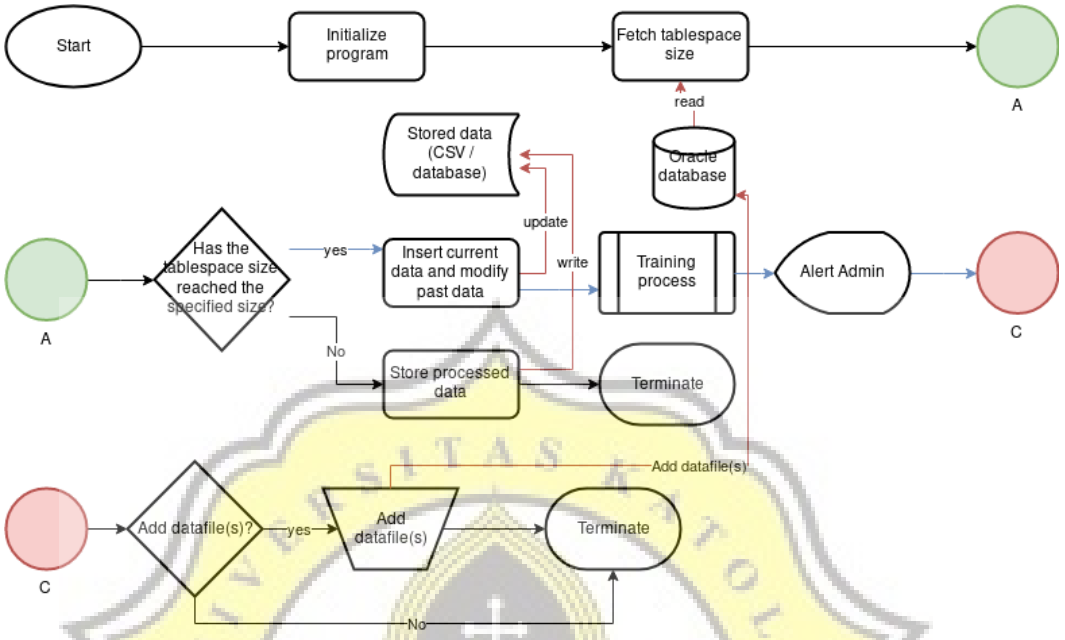

Illustration 4.5: Mine dataset process

Database administrator first need to initialize/run the program. This program launched by the cron job.

Then, after it got launched, get the current tablespace size (max size and current size) from the database.

After it read the size, check whether the tablespace has reached a specified size to consider it as “full”. This size get fetched from the settings file.

When it is not, process and store the processed data to data storage, then terminate the program. The crucial data that need to be stored are the current size (both maximum and current size), date when this data taken, and other that may be needed.

If the tablespace considered full, insert the current data, then modify the past data which doesn't have any output until this current date.

It also check the past data that if there's no empty output, get the latest row's output and set this current data's output as that latest row's

output. This serves as a method when the tablespace still being considered full but the database administrator doesn't add a new datafile.

What this output means is data type of date when will the cycle of this row of data being full. For example, the past data of 1 to 14 October doesn't have any output because the tablespace where this data is fetched from is still not full, because of that, when the data is full today (October 15th) make all of these past data's output to current date. This output will be used for back-propagation in the neural network because we will use supervised method. After that, continue to the training process (next section).

When the training process is finished, alert the database administrator that the current tablespace size has reached certain threshold. After administrator received alert, they can add datafile(s) to the database or not. This is done manually by the admin. Then after alerting the admin, terminate the program.

b) Training process

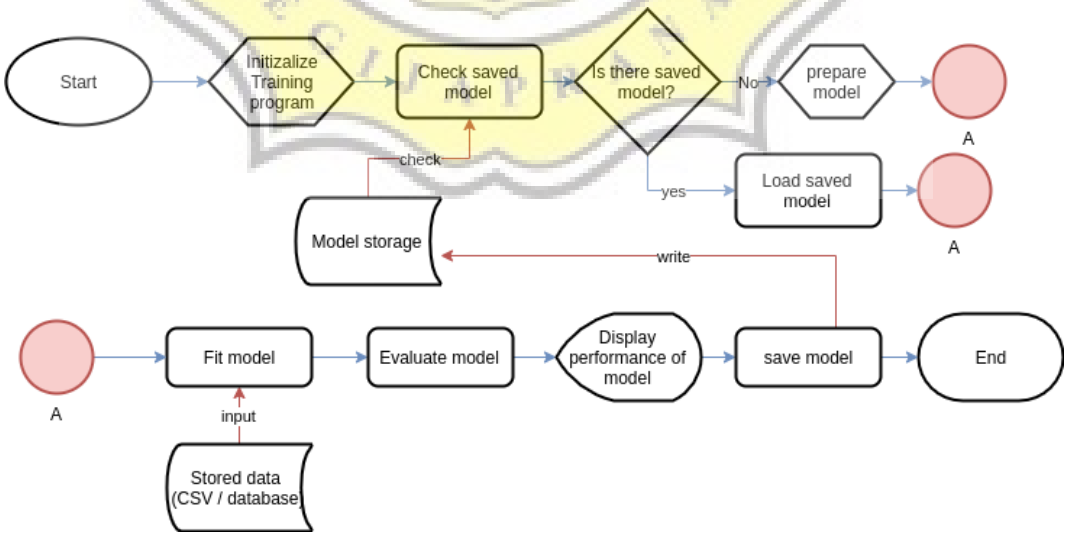

Illustration 4.6: Training Process

The task of this process is to train the neural network model. Then in the future, this model then will be used for predicting the time of when the database will be full, based on the current state of the tablespace.

This process runs only after it get invoked inside mine dataset process.

After this process got invoked, it first checks if a saved model exist.

If there's no saved model, prepare new model with set of input, hidden layer, output, bias, and weight. The amount of neurons in input layer, and hidden layer should be the same for the entire system at any point in time. When there's any change in the model for any point in time, the model need to be retrained from the start of the program, and it will cause a performance issue.

When saved model is found, load the model. This process will shorten the time of training because it only need to train from the current new data in the fit model process.

After prepare / load model, the process will be the same until it get terminated. First get the stored data, then fit the model. This fit task is training the data generated from mine dataset process. Neural network do training by calculating the input data against weight and bias in each connection of every layer. Then the output data is a value that's produced by neural network. This value is calculated by a function against the actual output of when the tablespace actually become full. Returned value from the function is called margin of error and this margin of error is used to do Back-propagation. Do this calculation and back-propagation using all data in the data storage or only the new data.

After fitting the model, show the performance of the current model. Then save the model to storage to use later for future training or predicting.

c) Predicting process

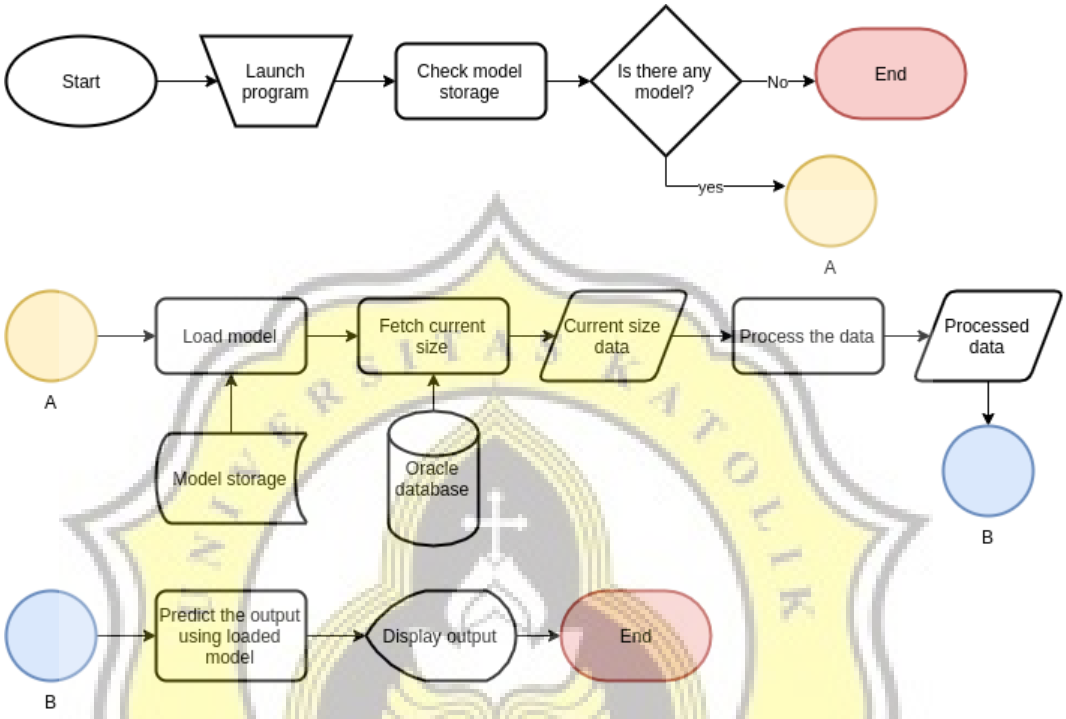

Illustration 4.7: Predicting Process

With this process, database administrator can predict when the database will be full by using trained model.

First administrator launch the predictor program, then it automatically checks the model storage if there's any model can be used. If there's no model that can't be used, terminate the program.

When it found any model, load that model, then fetch the current tablespace size from the oracle database. Process this data before predicting the output as an input. After the data being processed, predict this data using the loaded model.

The loaded model contains the number of input neuron, total neurons of hidden layer, output layer, weight and bias. The processed data contains data that will get fed into the machine learning. Data that are

crucial in this process are the size of the current tablespace and other data that may affect the output. This data is called "Input Data". The amount of input data should be the same as the amount of input neurons in the model. Total of input data should be the same when the model got trained. After it get fed into the input part of the neural network, it calculates from the current model (weight and bias) and it produces a value in the output layer. This value is the number of day(s) the tablespace will be full.

Finally, display the output of this prediction then terminate the program.

Database administrator only need to launch this program anytime then it automatically predicts the output using available model.

4) Manual and Automatic Part of the System

a) Database administrator need to setup an user in the oracle which has "*GRANT SELECT ON dba_free_space TO demouser;*" , "*GRANT SELECT ON dba_data_files TO demouser;*" and other privileges mentioned by settings program (create session, create tables, trigger, and others).

b) The system automatically train the neural network (training process) if the current tablespace size reach certain threshold.

c) Database administrator need to set the threshold of considered "full" tablespace size and the time that the program mines the dataset.

d) Database administrator must launch and run the mining process first in order for the system to work.

e) Database administrator can launch predict program anytime with the requirement of model saved.