# "WINSEER" FAILURE PREDICTION FOR DECISION MAKERS IN DATA CENTERS USING DATA MINING

D.G.S.M. Wijayarathne[1], W.K.S.D. Fernando[2], M.P.L. Mendis[3], J.S.D. Fernando[4], A.S.M.S Sharfaan[5] and C.D Manawadu[6]

[1] Faculty of information technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka, Email: samwijayarathne@gmail.com,
[2] Faculty of information technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka, Email: samithdf@gmail.com,
[3] Faculty of information technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka, Email: mplmendis@gmail.com,
[4] Faculty of information technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka, Email: shaminidhanushika@gmail.com,
[5] Faculty of information technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka, Email: saumy.saleem@hotmail.com,
[6] Faculty of information technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka, Email: dilhanmanawadu@gmail.com

## ABSTRACT

In the modern world organizations, data centers act a very important part and it is continuing to grow in their scale and complexity. Sometimes these data centers are prone to errors and face failures. These failures badly impact on organization and leads into massive troubles with data losses. Then the organizations have to replace, repair those data centers with minimum time consuming. Currently there are systems for data center monitoring, online repairing, failure prediction for data center administrators and more. Classical theories and conventional methods do rarely consider the actual state of a system and are therefore not capable to reflect the dynamics of runtime systems and failure processes. This project will present an unsupervised failure detection and prediction method using a data mining model. It characterizes the previous failure records of the system and predicts anomalous behaviors. Since the existing researches have only focused only on failure prediction, this research will lead to a unique area, data center failures prediction that addresses organizational decision makers to ease the organizational decision making process. This will help to shorten the lifespan of a failure recovering process in an organization. This will be to research community on further researches on increasing data center dependability. And for all kind of failure prediction researches in general.

***Keywords***: Data centres, failure detection, failure management, failure prediction, Decision makers, Data mining**.**

## 1. INTRODUCTION

In The modern world organizations, data centres act a very important part and it is continuing to grow in their scale and complexity. Each piece of these data centre infrastructure must operate efficiently to maintain high performance and availability. Conversely, any anomaly within these infrastructures can quickly degrade the user experience and business effectiveness.

Once captured a failure of a data center, this information is saved in a log file. These log files can be used to generate reports regarding the health of the servers. This process enables the end user to react to an event. But, it means that the damage has already been occurred, and the servers has crashed [1]. But these past failure reports also can be used to generate intelligent decisions about the health of the servers.

Data Centers, data center is a facility used to house computer systems and associated components, such as telecommunications and storage systems. It generally includes redundant or backup power supplies, redundant data communications connections, environmental controls and security devices. A wellmanaged data center ensures that business never suffers because of one failure somewhere. It also supports the growing needs of a business effectively. And also data center is a highly scalable entity in an organization. If there is any failure in data center servers, it will be a big disadvantage for the whole company and it will badly impact on organizational budget [2, 3]. Therefore, this is an important area to be concern.

The research group has collected information for this project, through related websites, referring books about data centers, data center failures, data mining algorithms, data ware houses, and so on. And also basically previous researches which consist of researches about data center failure monitoring, network failure monitoring. A team has builds a proactive prediction and control system for large clusters; some analyze the correlation between failures. That was much related to this problem and was a great help to get more information about this research.
The structure of this project proposal document is as follows;

Section 1 indicates the overview of the project and its problems related to the current situations faced by the companies or organizations. Section 2 describes the literature review in detail with research gap. Section 3 describes the objectives of the project, section 4 describes the methodology, the team use to develop this project and it indicates the flow of the project. Section 5 describes the budget of the project.

## 2. LITERATURE REVIEW

Similar study was conducted by Guan et.al, who presented an unsupervised failure detection and prediction procedure using an ensemble of Bayesian models. They tackled the problem from an anomaly detection viewpoint. It works in an unsupervised learning manner and deals with unlabeled datasets. This model estimates the probability distribution of runtime performance data collected by health monitoring tools when servers perform normally. Apart from Bayesian models, they used decision trees and presented a failure prediction method using that. They implemented a prototype of failure detection and prediction mechanism and evaluate its performance on a data center test platform. Experimental results show that their proposed manner can predict failure dynamics with higher accuracy. In this paper they mainly focus on administrative mistakes due to failures of the complex data centers [4].

Sahoo et .al. attempted to build a proactive prediction and control system for large clusters. They collected event logs containing various system reliability, availability and serviceability (RAS) events, and system activity reports from a 350-node cluster system for a period of one year. They applied a faltering technique and model the data into a set of primary and derived variables. They had applied time-series algorithms, rule-based classification techniques and Baysian network models to predict failure in a cluster [5].

Yamanishi et.al. introduced a new methodology of dynamic syslog mining for network failure monitoring. Key features of that model are I) probabilistic modeling using an HMM mixture, II) online discounting learning of parameters in the model, III) dynamic model selection for determining the optimal number of mixture components and IV) scoring sessions using universal test statistics with a dynamically optimized threshold. Detection of failures or their indications from syslog can be reduced to the issue of anomalous session discovering. Their process can be directly applied to the analysis of a vast range of event log files, including system calls, command lines and web access logs. In this paper they have mainly concentrate on mining symbolic data. [6].

Dudko et.al. described state-of-the art failure monitoring and prediction research, the short- comings of current models in the context of frameworks such as Hadoop, and propose a novel approach to predict performance and failures in Hadoop clusters. Although they presented a research in the context of Hadoop systems, idea is generalizable to similar cloud frameworks [7].

Patnaik et.al. presented a temporal data mining solution to model and optimize performance of data center chillers, a key component of the cooling infrastructure. It helps bridge raw, numeric, timeseries information from sensor streams toward higher level characterizations of chiller behavior, suit- able for a data center engineer. To aid in this transduction, temporal data streams are first encoded into a symbolic representation, next run-length encoded segments are mined to form frequent motifs in time series, and finally these metrics are evaluated by their contributions to sustainability. A key innovation in their application is the ability to intersperse ‘do not care’ transitions (e.g., transients) in continuous-valued time series data, an advantage they inherited by the application of frequent episode mining to symbolized representations of numeric time series. Their approach provides both qualitative and quantitative characterizations of the sensor streams to the data centre engineer, to aid him in tuning chiller operating characteristics. This system is currently being prototyped for a data center managed by HP and experimental results from this application reveal the promise of our approach [8].

Mickens et.al introduced new techniques for predicting availability and tests them using traces taken from three distributed systems and described three applications of availability prediction. The first, availability-guided replica placement reduces object copying in a distributed data store while increasing data availability. The second shows how availability prediction can improve routing in delay-tolerant networks. The third combines availability prediction with virus modeling to improve forecasts of global infection dynamics [9].

## 3. OBJECTIVES

This ‘WinSeer’ project is mainly aimed on the decision makers of the particular company or organizations. Therefore this research is about to predict about data centers’ failures through data mining. In order to predict the failures of data centers, team members analyzed the data collected from previous records of failures. Therefore this ‘WinSeer’ monitor the data center failures and inform it to decision makers while keeping the data center alive.

- Select best data mining algorithm.
- Collect historical data set.
- Develop a data mining model.
- Train selected algorithm according to the data set.
- Acknowledge decision makers about failures.
- Testing the WinSeer Project.

## 4. METHODOLOGY

Data centre is basically a storage centre of all servers and related hardware equipment. A well-managed data centre ensures that business never suffers because of one failure somewhere. It also supports the growing needs of a business effectively. A data centre is a highly scalable entity in an organization. If there is any failure in data centre servers, it will be a big disadvantage for the whole company. For an organization increasing expenditures is a huge thing to be concern. There are systems to notify the failures of data centre servers, but there is no such a system to predict about a failure in a server. The Research was mainly aim the Failure Prediction for decision makers in Data centres using Data Mining. After that the team has decided on what type of application that the team will going to create and just inspected how the process are conducted according to the requirements and the data have gathered [10, 11].

**Planning**

This phase is the most critical and important part in the software development life cycle (SDLC). In this phase project team identified why an information system should be built and determined how the project team will build it. The team decided to build this failure prediction system because there is no any system to predict the failures in data centers.

Therefore the team has identified business value of this system, analyzed the feasibility, developed a work plan, and staff the project and control and direct project.

**Analysis**

The analysis phase answers the questions of who will use the system, what the system will do, and where and when it will be used in SDLC. The team needed information about data centres' failures, suitable data mining algorithms, dataset information, device previous failure information and so on. Team has collected this information through questionnaires, interviews with project managers, data center managers as well as the decision makers and other relevant people in different companies in Sri Lanka. And also team used books, internet articles and so on. Since this project was based on data centres, information gathering about data centres' failures to be an essential part. Therefore the team has got data centres' information by contacting the project managers of several companies.

Then can get the clear idea about data centres' failures. After gathering all requirements, team analyzed problems and designed it in a graphical way to get a good idea about the requirements and the application need to implement. It was very significant to understand the problem before designing and the implementing the solution. By using this way, team took an idea about user's acceptations and what would they need. And it was very easy to side track the errors and ensured all the requirements are completed [11, 12].

**Design**

The design phase describes how the system will operate, in terms of hardware, software and network infrastructure. The primary objective of the design phase is to create a design that satisfies the agreed application requirements. The team has developed a system to predict data centres' failures early hand and inform in to decision makers. Activities mainly grouped in to two categories.

1. High level design (Architectural)
This will deliver interfaces and the relationship between them. The outcome will be a structure charts or a software architecture.

2. Low level Design (Detail)
This will be the detail specification of interfaces. Object oriented designing diagrams used for each module. Team used UML to design diagrams. Such as Use Case diagram, Class diagrams, Sequential diagram, etc.

**Implementation**

Implementation is the final stage of software development life cycle. Importance of this phase is very high comparing to the requirement gathering, requirement specification and designing.

**Construction**

The development team needed previous data records of a particular organization to create the data mining algorithm. Therefore team got an opportunity to get past two years data records from a recognized company. It is an IT company which has worldwide connections in IT field. The team got a large amount of data set and it is around 150 million records. Therefore it was difficult to handle and also it contains a lot of unusable records. The original data set that the team got from the company was in comma separated value format and first it was imported to the SQL database. It gave a database with the size around 30 Gigabytes. Then team has analyzed the data set and removed some unnecessary records and processed the dataset with the SQL server. To feed the data into WEKA, team has to put it into a particular format.

WEKA's preferred method for loading data is in the Attribute-Relation File Format (ARFF), where can define the type of data being loaded, and then supply the data itself. In the file, team has defined each column and what each column contains.

## 5. Results and Discussion

'WinSeer' was a system uses to get predictions about failures in data centres' servers before a failure happens. This will be more useful for the decision makers who are in an organization. Because they will get an alert messages about device failures beforehand and can prevent disasters occur by failures. The target of the team was to develop a system with high accurate prediction capabilities. In order to get a better result from the system team compared several algorithms before creating the 'WinSeer' system.

Those algorithms were Linear Regression, Clustering, Nearest Neighbour IBK, M5P, Decision Stump, SimpleKmean, RepTree, Decision Tree-J48 Draft and Decision Tree-J48. By comparing accuracy of those algorithms, and the mining models created with each, team decided Decision Tree-J48 was the most appropriate algorithm to create the 'WinSeer' mining model. Since its' given the highest accuracy rate.

By using the 'WinSeer' system, decision makers in an organization can get alert reports according to the system generated predictions and the current status of the devices. And also can get e-mail alerts about critical devices in their organization. Therefore decision makers can get immediate actions to prevent those failures. It will increase their economy, save time and can create good image of the company.

## 6. CONCLUSION

**Importance of Outcome**

In the modern world organizations data centres are prone to errors and face failures. These failures badly impact on organization economy as well as organization's image and these failures leads into massive troubles with data losses. In large-scale and complex data centres are susceptible to hardware devices failures and decision maker's mistakes, which significantly affect the system performance and management. Therefore 'WinSeer' system is very important for the decision makers in order to take their decisions more accurately, timely and more meaningfully. By using this system before a failure occurred, decision makers will get alert notifications about servers which are going to be fail. Predict of such failure can lead to greater efficiency of both people and equipment. This will save time of the internal staff and this will increase the organizational economy. It will leads to more reliable operation of data centre servers and reduce overall time to repair a problem and it will increase the flexibility and dramatically lower cost of new servers as well as it will be good advantage for business and for decision makers. Therefore organization can build up an attractive image between both internal staff and their customers.

**Limitations**

- It is not possible to import more than 4.0GB data in to MS SQL at once.
- Lack of reliable data limit the scope of project analysis.
- Accessed are denied for other adequate, timely data resources.
- Time management is a limitation for a good research project.
- RAM limitations are a problem for the continuation of the project.
- Lack of experts for editing and proofreading within the group.
- Study effort needed for the new software and features.
- Lack of available tutorials and publications about Weka data mining tool.

**Future Work**

In the future, 'WinSeer' system could be upgraded with the following features:

- Integrate with mobile application and send an alert to the decision maker when a 'warning' prediction is generated.
- Build a component which can be handling by decision makers and send messages to relevant people to take actions immediately as soon after he got the alerts.
- This 'WinSeer' system can merge with other devices (apart from data centres) which have past failure records and can get failure prediction.

## 7. ACKNOWLEDGEMENT

Our sincere gratitude goes out to all the people who have contributed to the success of this research paper specially the team from Zone24x7 Inc. And finally we thank all the anonymous reviewers for their insightful comments.

## 8. REFERENCES

[1] HowStuffWorks.com Contributors, "Are data mining and data warehousing related?", 20 April 2011. *HowStuffWorks.com*, [Online]. Available: http://www.howstuffworks.com/are-datamining-and-data-warehousing-related.htm. [Accessed: March. 23, 2013].

[2] "Database Fundamentals," 2008. [Online]. Available: http://www.personal.psu.edu/glh10/ist110/topic/topic07/topic0 7_09.html. [Accessed: Mar. 23, 2013].

[3] B. Sudeshna, Georgia, "DATA MINING," 1997. [Online]. Available: http://www.siggraph.org/education/materials/HyperVis/applic at/data_mining/data_mining.html [Accessed: Mar.23, 2013].

[4] R. K. Sahoo, A. J. Oliner, I. Rish, M. Gupta, J. E. Moreira, S. Ma, R.Vilalta, and A. Sivasubramaniam, "Critical event prediction for proactive management in large-scale computer clusters," In Proceedings of ACM International Conference on Knowledge Discovery and Data Dining (KDD), 2003.

[5] K. Yamanishi and Y. Maruyama, "*Dynamic Syslog Mining for Network Failure Monitoring"*, KDD'05, 2005.

[6] Q. Guan and S. Fu, "auto-AID: A data mining framework for autonomic anomaly identification in networked computer systems," In *Proceedings of IEEE International Performance Computing and Communications Conference (IPCCC),* 2010.

[7] D.Patnaiky, M.Marwah, R.Sharma and N.Ramakrishnan, “*Sustainable Operation and Management of Data Center Chillers using Temporal Data Mining*”, pp. 1305-1314,In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*

[8] J. W. Mickens and B. D. Noble, “Exploiting availability prediction in distributed systems,” In *Proceedings of USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2006.

[9] umsl.edu. "*Prototyping in Systems Analysis*.". [Online]. Available.
http://www.umsl.edu/~sauterv/analysis/488_f01_papers/Hammer/term_paper_body.htm#Introduction.
[Accessed: Feb.13, 2013].

[10] “Data center management,” 2013. [Online]. Available:
http://www.webopedia.com/TERM/D/data_center_manageme nt.html. [Accessed: Feb 20, 2013].

[11] “Software Development Life Cycle (SDLC) Phases,” NOVEMBER 27, 2011. [Online]. Available:
http://www.sdlc.ws/software-development-life-cycle-sdlcphases/. [Accessed: Feb.13, 2013].

[12] DOJ System Development Life Cycle Guidance, "Implementation Stage of SDLC," November 2005. [Online]. Available:
http://www.writework.com/essay/implementationstage-sdlc. [Accessed: Feb.13, 2013].