Computers and Electrical Engineering 71 (2018) 93–101

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Real-time big data analytics for hard disk drive predictive maintenance[☆]

Chuan-Jun Su, Shi-Feng Huang[*]

Department of Industrial Engineering & Management, Yuan Ze University, 135, Far-East Rd., Chung-Li, Taiwan, ROC

ARTICLE INFO

Keywords:
Big data
Predictive maintenance
Hard Disk Drive (HDD)
Apache Spark
Machine learning

ABSTRACT

Effective and reliable cloud services depend on the quality of service provided by large-scale data centers, and data center equipment reliability issues can cause significant data and financial loss to cloud service clients. Corrective maintenance is a reactive approach that only corrects problems once they occur, resulting in unwanted downtime, while preventative maintenance relies on replacing equipment which may yet still have considerable effective operating lifetime, thus raising maintenance costs. In contrast, predictive maintenance can potentially predict equipment failure in advance, thus reducing unplanned downtime and extending equipment lifetimes, thus reducing maintenance costs while increasing system reliability. This research aims to develop a real-time predictive maintenance system, HDPass, based on Apache Spark for the detection of imminent hard disk drive (HDD) failures in data centers.

1. Introduction

Cloud computing and virtualization technologies have emerged as a compelling paradigm for computing at scale and the delivery of online services, turning data centers into an indispensable infrastructure for service providers to furnish continuous and diverse services over the Internet. Highly reliable, uninterrupted data storage is a critical element of this infrastructure [1]. Hard disk drive (HDD) failure poses a constant threat for potentially financially devastating loss of customer information, business transactions, and sales data. Hard disk manufacturers claim failure rates are below one percent, but independent studies have shown that failure rates can frequently reach 14% [2]. Even a one percent failure rate is a serious problem because disk failure often occurs without warning, raising the urgent need for a means of predictive maintenance for HDD.

“Preventive Maintenance” seeks to prevent equipment failures through the use of a pre-determined schedule of planned maintenance actions based on time passed or meter triggers. Preventive maintenance routines are frequently regardless of actual device condition, and could give managers a false sense of security and in fact end up wasting valuable resources. Predictive Maintenance (PM) is a condition-based strategy that predicts machine failures in advance based on historical data, and this approach is not only more cost-effective but also effectively increases equipment life cycles.

Big data analysis techniques allow for the systematic analysis to reveal hidden patterns in huge streaming data sets. This has facilitated the development of preventative maintenance (PM) and its use in a wide range of monitoring applications, such as vibrations, temperature, and noise. In the field of computer management and intelligence, log file analysis is a common approach to monitor system status. Laprie defines system failure as “… when the delivered service deviates from the specified service, where the

[☆] Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. J. S. Sheu.

[*] Corresponding author.

E-mail addresses: iecjsu@saturn.yzu.edu.tw (C.-J. Su), iesfhuang@saturn.yzu.edu.tw (S.-F. Huang).

https://doi.org/10.1016/j.compeleceng.2018.07.025
Received 15 September 2017; Received in revised form 17 July 2018; Accepted 24 July 2018

service specification is an agreed description of the expected service." [3] Despite being inherently prone to failure, computer systems are increasingly essential to nearly all aspects of human life and temporary service failure can have expensive and potentially life-threatening consequences. Failure prediction can be performed by analyzing the system logs to identify indicators of an imminent malfunction. Many studies have investigated how to improve the performance of proactive fault management for computer systems [4].

In this paper, we describe the development of a real-time prediction system that can assist IT teams in maintaining large scale storage systems by issuing warnings for imminent drive failure. We also describe a framework for the predictive monitoring of hard disk drives (HDD) failure by analyzing machine log files rather than using traditional statistical prediction methods. The development process involves the following tasks: (1) Retrieve and pre-process HDD data; (2) train and verify a prediction model; (3) build a prediction system based on the model; and (4) predict HD failures in real-time based on streaming data.

1.1. Predictive maintenance in big data analysis

A failure precursor is an event or series of events that is indicative of an impending failure [5]. A failure can be predicted by correlating changes in monitored precursor parameters. By identifying the precursor parameters and monitoring them, the extent of deviation or degradation from expected normal operating conditions can be assessed. This information can be used to provide advanced warning of failures and for improving qualification of products. Zhang [6] presented a Prognostics and Health Management (PHM) approach for power supplies. He firstly analyzed historical data to identify precursor parameters, then conducted experiments under different environmental and usage conditions to establish a baseline, followed by precursor parameter identification for one switch-mode power supply.

PM emphasizes the direct monitoring of equipment performance during normal operations to anticipate future failure. Rather than scheduling equipment maintenance after a certain number of operating hours regardless of performance, data on vibration, temperature, and other variables can be collected to predict imminent failure through modeling and analysis. Big data analysis techniques can process large-scale data flows streaming from multiple sources in real-time, allowing PM to perform effective diagnosis and maintain large numbers of devices while minimizing equipment maintenance costs. Furthermore, big data techniques allow PM to continuously collect and analyze data to identify patterns for the continuous improvement of device performance. Edge [7] proposed a fraud management and a fraud prevention architecture based on implementing fraud policies, which can process real-time data streams, proactively detect fraud and block suspicious transactions. Lian [8] studied the problem of matching both static and dynamic patterns over streaming time-series data in stock data monitoring and proposed a novel multistep filtering mechanism. Ko [9] proposed a novel measure for intrusion detection and surveillance using pure time series data streams. Yang [10] used the Bayesian robust principal component analysis (RPCA) approach to develop a new method for detecting traffic events that impact road traffic conditions. This method processes data streams incrementally with small computational costs, and can detect real-time online events.

1.2. Self-monitoring and reporting technology (SMART)

All hard disk drives (HDD) and solid-state drives (SSD) encapsulate SMART data that depicts internal information about the drive. Disk monitoring is currently performed mainly through threshold analysis, which is predefined by disk manufacturers. In this approach, the device firmware compares the thresholds with the measured parameters; if a SMART attribute drops below its threshold it indicates a potential problem with the drive.

Some studies have attempted to improve the threshold methodology by exploring critical factors for drive failure. Hamerly proposed two Bayesian methods to predict disk drive failures based on measurements of internal drive conditions [11]. They performed sensitivity and specificity analysis on a Quantum SMART dataset, with results showing that drive failure is positively correlated with power-on hours, read error rate and spin-up time. Hughes proposed improved methods for disk-drive failure prediction [12], using SMART warnings for HDD failure to analyze parameter value intervals to enhance disk-drive failure detection accuracy. Agarwal constructed a decision support system to detect drive failures and analyze their cause [13]. He found that the accuracy of disk-drive failure prediction can be improved by exploring rules from disk events and developing a black-box model.

1.3. Machine learning

Data mining typically uses machine learning techniques for prediction or classification. With machine learning, the computer makes a prediction and then "learns" from subsequent feedback, including examples and domain knowledge. When a similar situation arises in the future, this feedback is used to make the same prediction or to make a completely different prediction. Sample data is used to train the machine learning system to properly perform the desired task and product a model which is then applied to perform predictions on new datasets.

The proposed HDPass system can learn failure patterns from historical data, allowing it to subsequently predict future breakdowns of targeted devices. The core method of HDPass is ensemble learning which improves learning results by integrating multiple algorithms to achieve better performance than could be achieved from any constituent learning algorithm alone.

Ensemble learning has several classifiers, such as Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Random Forest (RF). The performance of Random Forest compares very favorably with that of other classification algorithms and has been widely applied for ensemble learning. Compared with the global machine learning models, such as ANN and SVM, which try to build