Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 00 (2018) 000–000

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA

Moving towards Real-time Data-driven Quality Monitoring: A Case Study of Hard Disk Drives

Ardeshir Raihanian Mashhadi[a], Willie Cade[b], Sara Behdad[a,c]*

[a]*Mechanical and Aerospace Engineering Department, University at Buffalo, Buffalo, NY, 14260, USA*
[b]*ICR Management,, Chicago, IL, 60651, USA*
[c]*Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, NY, 14260, USA*

* Corresponding author. Tel.: +1-716-645-5914; fax: +1-716-645-2883.
*E-mail address:*sarabehd@buffalo.edu

Abstract

Since its emergence, the cloud manufacturing concept has been transforming the manufacturing and remanufacturing industry into a big data and service-oriented environment. The aggressive push toward data collection in cloud-based and cyber-physical systems provides both challenges and opportunities for predictive analytics. One of the key applications of predictive analytics in such domains is predictive quality management that aims to fully exploit the potentials provided by the enormous data collected via cloud-based systems. As a case study, a data set of hard disk drives' Self-Monitoring, Analysis and Reporting Technology (SMART) attributes from a cloud-storage service provider has been analyzed to derive some insights about the challenges and opportunities of using product lifecycle data. An analysis of time-to-failure monitoring of hard disk drives in real-time has been carried out and the corresponding challenges have been discussed.

© 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of NAMRI/SME.

Keywords: Cloud manufacturing; data-driven quality and health monitoring; predictive maintenance; hard disk drive failure prediction

1. Introduction

Since the emergence of cloud manufacturing [1], there has been an aggressive push toward data collection for predictive analytics purposes [1,2]. The transition from product-based manufacturing systems to service-oriented platforms paves the way for solving the informational and data bottlenecks in manufacturing applications [3]. Among such applications, predictive health management has been introduced as a key enabler of future manufacturing systems within cyber-physical platforms [4]. Capabilities provided by the Big Data environment allow various sorts of analytical, statistical or data-driven predictive

2351-9789 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of NAMRI/SME.

algorithms [5] to be applied to manufacturing problems in order to improve performance. One of the key manufacturing concerns relates to quality control and health monitoring. Utilizing Big Data capabilities helps to ameliorate such concerns by applying statistical and data-driven algorithms on the data gathered from the manufacturing equipment or the final products. For example, despite the relatively young age of such concepts, some applications in health assessment and prognostics [6], machine availability monitoring [7] and machine tool monitoring [8] have been introduced and developed based on the Big Data capabilities and advanced analytics.

Cloud-based capabilities are changing the way data collection and the corresponding analyses are carried out for quality control and health management. Table 1 compares some key features of quality control in traditional and cloud-based manufacturing paradigms. While the transition toward cloud-based quality control and health monitoring can mitigate the cost of data collection, the enormous amount of data generated every day by each manufacturing and remanufacturing plant would make data storage, warehousing and computation more challenging and costly. On the other hand, compared to the traditional quality control methods that aim at collecting specific performance measures data, cloud-based platforms need to handle the real-time analysis of heterogeneous data types and complex working conditions. Challenges related to heterogeneous data formats and complex working conditions in cloud-based machine health monitoring have been previously highlighted [9].

Table 1 – Comparison of data collection and analysis for health management between traditional and cloud-based systems

	Traditional	Cloud-based
Time interval	Less frequent	Very frequent
Type of data	Specific	Heterogeneous data types
Data analysis	Offline	Real-time
Data collection cost	High	Relatively lower
Data storage and warehousing cost	Low	High

This paper provides a case study of cloud-based product monitoring in real-time. A massive data set of hard disk drives' Self-Monitoring, Analysis and Reporting Technology (SMART) attributes from a cloud storage service provider has been analyzed and a real-time quality monitoring scheme has been provided. Furthermore, the potential challenges that need to be overcome in such systems have been discussed.

The rest of this paper is structured as follows. Section 2 provides a brief background of the hard disk drive failure prediction studies, illustrating the gap that could be covered via cloud-based systems. Section 3 provides the analysis of the data of a cloud-storage provider and discusses the potential challenges that should be addressed. Finally, Section 4 concludes the paper.

2. Background

Most of the data produced in the world are stored on hard disk drives [10], therefore, their imminent failure prediction has been a subject of considerable importance in the literature. Previous efforts have been made to use SMART attributes, which are real-time measurements of the drives' technical status, in order to predict hard drive failures.

Older failure prediction studies focused on critical thresholds on SMART attributes for failure prediction [11]. Machine learning algorithms and statistical tests have been utilized subsequently for disk drive failure predictions [12,13]. For example, Zhu et al. [14] have investigated the usage of a Support Vector Machine (SVM) model and backpropagation neural network models on a dataset of 23395 drives for failure prediction. They argued that while SVM provides the lowest false alarm rate, the neural network model presents a better failure detection rate up to 95%. However, Li et al. [15,16] highlighted the fact that the accuracy of the failure prediction models may not reflect their practicality. They proposed two different metrics for disk failure based on the probability of data being at risk. They used classification trees and recurrent neural networks for failure prediction and gradient boosted regression trees for residual life prediction. Similarly, Pang et al. [17] developed ensemble classification models to predict various levels of the remaining working time of the drive. However, most of the studies focus on the benefits of using SMART attributes for homogenous drive populations. Rincon et al. [18] have recently extended such analyses to more heterogeneous data from data centres. They used decision tree, neural network, and

logistic regression.

While the previous efforts have mostly focused on the prediction of an impending failure or the possibility of a failure within a fixed horizon, such efforts could be improved by providing a real-time continuous quality monitoring measure via the data provided by the cloud services. From the cloud manufacturing standpoint, investigating hard disk drive data is substantial, as these devices are among the first mediums that were utilized in cloud-based service-oriented platforms. In order to investigate the potentials provided by the cyber-physical systems regarding quality monitoring, we present a case study of hard disk drives' time-to-failure prediction. However, the framework presented and the challenges mentioned can be applied to any other cloud manufacturing and remanufacturing setting.

3. Case study: hard disk drive time to failure prediction

3.1. Datasets

In order to investigate predicting hard disk drives' time to failure using SMART attributes, two datasets have been analyzed. The datasets have been generated based on the Backblaze raw hard drive test data [19].

The first data set-*'A'*- consists of SMART stats for 91,701 drives, including 6854 failed drives. The Backblaze data report the SMART values of hard drives at the end of each day in the period of 2013-2017. The Backblaze data report 40 different SMART stats for data related to 2013-2014 and 45 SMART stats for data related to 2015-2017. For both data sets, the columns that contained missing data in more than one-third of the rows have been removed.

Table 2 summarizes the SMART attributes and their definitions that have been considered. The data include a failure identifier that indicates whether or not the hard drive was in working condition at the end of the reporting day. The drives that failed were replaced the next day and would be removed from the data sets of the subsequent days. Therefore, the failure dates of the failed drives can be extracted from the Backblaze data. The second data set-*'B'*- contains a random sample of 3297 hard drives' SMART stats and the corresponding time to failures.

Table 2 – SMART attributes and their definition

SMART Attribute	Definition
SMART 1	Read Error Rate
SMART 3	Spin Up Time
SMART 4	Start/Stop Count
SMART 5	Reallocated Sectors Count
SMART 7	Seek Error Rate
SMART 9	Power-On Hours
SMART 10	Spin Retry Count
SMART 12	Power Cycle Count
SMART 184	End-to-End error / IOEDC
SMART 187	Reported Uncorrectable Errors
SMART 188	Command Timeout
SMART 189	High Fly Writes
SMART 190	Temperature Difference or Airflow Temperature
SMART 191	G-sense Error Rate
SMART 192	Power-off Retract Count
SMART 193	Load Cycle Count
SMART 194	Temperature
SMART 197	Current Pending Sector Count
SMART 198	Uncorrectable Sector Count
SMART 199	UltraDMA CRC Error Count
SMART 240	Head Flying Hours
SMART 241	Total LBAs Written
SMART 242	Total LBAs Read

3.2. Analysis

Dataset *'A'* has been preprocessed and the columns containing missing values in more than one-third of the rows have been removed. Table 3 provides a comparison between the descriptive statistics of the failed and working drives' SMART stats.

The general procedure of the analysis is as follows. First, a statistical analysis is carried out on data set *'A'*. The purpose of this step is to evaluate the differences in values and trends of the SMART attributes across the failed and working hard disk drives. The SMART attributes that undergo great changes in drives that approach an impending failure may be good indicators for failure prediction. The next step is motivated by the mentioned point and is designed to look into the

relationship between SMART stats and the time to failure of the drives via regression analysis. In this step the SMART readings from dataset *'B'* are used as inputs of the analysis and the time to failure is the target value of the prediction. Figures 1 and 2 illustrate the trend of several common SMART attributes over time, up to the drive's failure date, for two different drives with significantly different life spans.

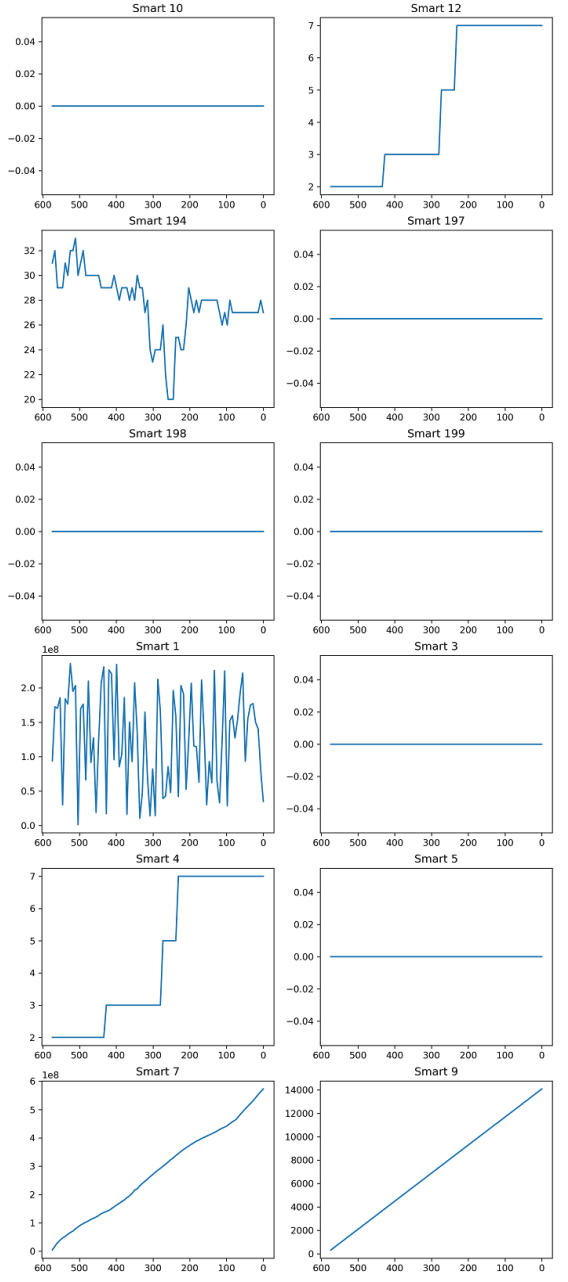

Figure 1 – Trend of SMART stats vs. time to failure for a drive with 600 working days

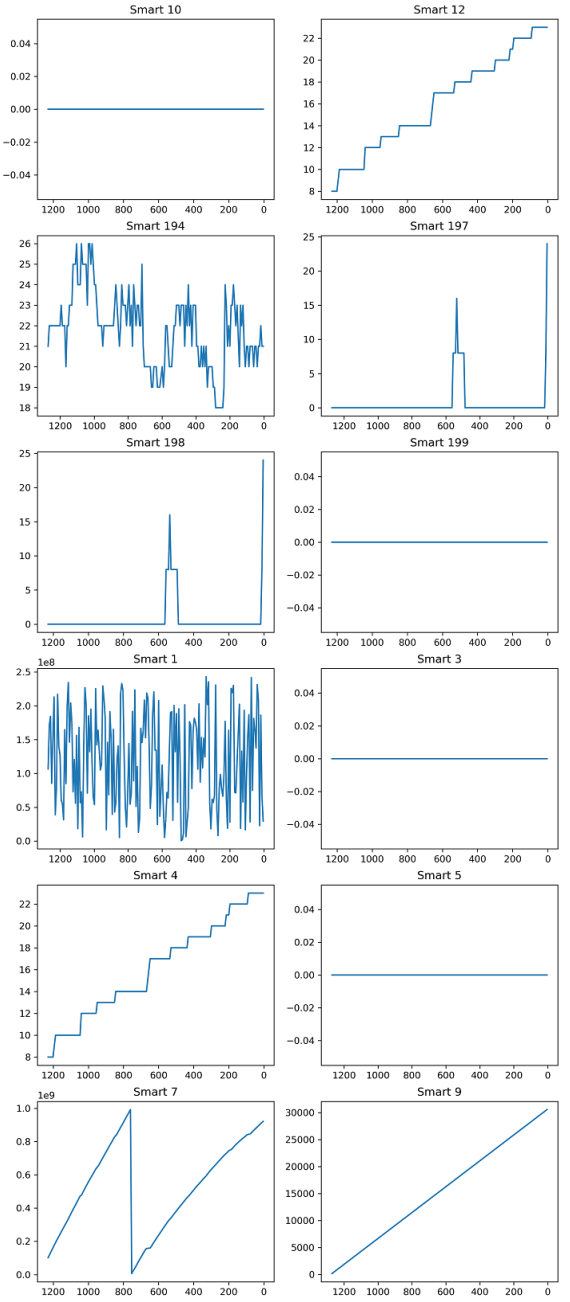

Figure 2 – Figure 1 – Trend of SMART stats vs. time to failure for a drive with 1200 working days

Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 00 (2018) 000–000

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

Table 3 - Comparison of descriptive statistics between the failed and working drives' SMART stats (dataset *'A'*)

	Mean		Std		Min		Max	
	Failed	Working	Failed	Working	Failed	Working	Failed	Working
failure	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00
smart_10	1.99E+03	2.62E+01	4.76E+04	2.03E+03	0.00E+00	0.00E+00	1.64E+06	2.62E+05
smart_12	23.34	8.80	28.77	58.42	0.00	0.00	1053.00	16419.00
smart_194	25.78	27.87	4.70	6.61	13.00	13.00	50.00	52.00
smart_197	301.15	0.04	3465.90	1.09	0.00	0.00	65534.00	80.00
smart_198	226.39	0.04	2968.88	1.07	0.00	0.00	65534.00	80.00
smart_199	5.38E+01	1.38E+01	6.51E+02	1.52E+03	0.00E+00	0.00E+00	1.98E+04	4.03E+05
smart_1	1.11E+09	7.56E+07	8.18E+10	8.10E+07	0.00E+00	0.00E+00	6.63E+12	2.44E+08
smart_3	402.26	249.79	1372.26	883.76	0.00	0.00	9370.00	9766.00
smart_4	36.25	10.66	468.61	79.59	1.00	1.00	26675.00	7796.00
smart_5	2117.90	2.03	8753.74	177.73	0.00	0.00	65224.00	39856.00
smart_7	1.51E+11	1.44E+10	3.99E+12	1.81E+12	0.00E+00	0.00E+00	2.19E+14	2.81E+14
smart_9	1.81E+04	1.63E+04	1.23E+04	1.10E+04	0.00E+00	0.00E+00	1.41E+05	6.57E+04

Table 4 – Descriptive statistics of the time to failures of drives in data set *'B'*

	Count	Mean	Std	Min	25%	50%	75%	Max
Time to failure	3297	329.41	269.88	0	109	262	493	1455

As can be seen, SMART attributes 3, 5, 10 and 199 seem to be stable and constant over time for both drives. SMART attributes 9 and 12 that respectively refer to power-on hours and power-on cycle counts, intuitively, increase over time. However, the difference between the step-like shapes of the SMART 12 trends for two drives suggests that the two drives underwent different power-on cycles. The rest of the attributes, while maintaining similar trends, seem different for each drive. These trends may be used for the remaining useful life predictions.

As can be seen, significant differences can be observed among the SMART stats of the failed and working devices. Higher values of read error rate (*SMART 1*), spin retry count (*SMART 10*), pending sector count (*SMART 197*), uncorrectable sector count (*SMART 198*), CRC error count (*SMART 199*), spin up time (*SMART 3*), star/stop count (*SMART 4*), reallocated sector counts (*SMART 5*) and seek error

2351-9789 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of NAMRI/SME.

rates (*SMART 7*) are observed in drives with an impending failure. Also, failed drives, on average, have higher power-on time and power-on cycle counts (*SMART 9 and 12*). Such differences may suggest a relationship between the SMART stat values and the time to failure.

In order to investigate the predictability of the time to failure based on the SMART values, regression analyses have been carried out on dataset *'B'*. Dataset *'B'* contains a random selection of SMART values and the corresponding time to failures at that point, over the lifespan of 3297 failed drives. In other words, Dataset *'B'* has been generated by sampling the SMART stats of the 6854 failed drives during their working days. Therefore, since the actual failure date of those drives is known, the time to failure can be calculated using the date of the SMART stat reading. Thus 3297 sets of hard disk drive SMART stats and the time to failure for the same drives are pooled. Figure 3 illustrates the distribution of the hard disk drive time to failures in dataset *'B'*.

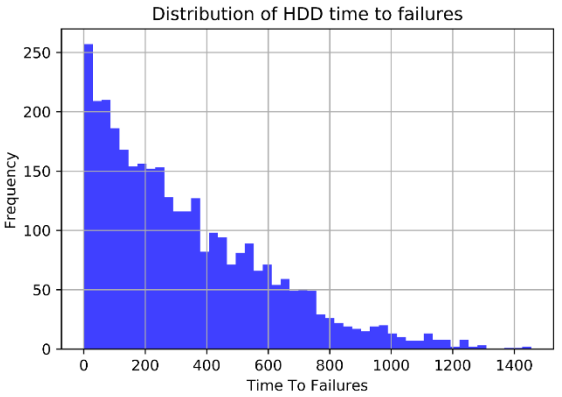

Figure 3 – Histogram of hard disk drives' time to failures in data set *'B'*

The time to failure values will be the output of the regression model and the SMART stats will act as the input variables. In addition, Table 4 presents the descriptive statistics of the time to failure values.

Figure 4 depicts the result of the dimensionality reduction applied to the SMART attributes in data set *'B'*. t-distributed Stochastic Neighbor Embedding (t-SNE) [20] algorithm has been used. t-SNE is a dimensionality reduction method that maps every data point to a location in a low dimensional space and has found to be more effective in data visualization compared to other high dimensional data visualization algorithms [20].

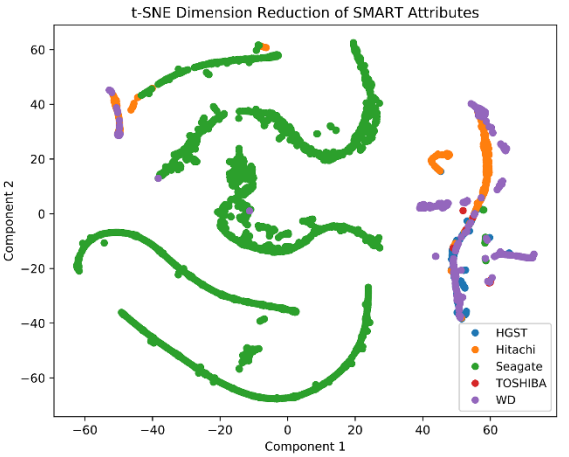

Figure 4 – Visualization of SMART attributes in data set *'B'*

The data points have been colored based on their brand. Note that the brand information has not been seen by the learning algorithm for dimensionality reduction and the color coding has been done after the neighbor embedding process. It is observed that a significant separation exists in SMART values of at least one brand compared to the rest of the brands. In other words, the data points of four of the hard drive brands cluster together relatively well and are separated from the fifth one. This highlights the challenges in failure prediction and time to failure estimation of hard drives based on SMART attributes in heterogeneous populations of drives. Based on this observation, three different regression models have been developed based on the brand information. One regression analysis has been done on the full data set *'B'* and two separate regression analyses have been done only on the two most frequent hard drive brands in the data set. The SMART attributes have been used as the predictor matrix and the time to failure has been considered as the output variable. Random forests [21] have been used as the regression algorithms. Decision tree models have been previously used for hard disk drive failure prediction. Moreover, ensemble decision tree models have been shown to be reliable and stable in failure prediction [15,16]. Random forests overcome the overfitting problem of the decision trees and also provide estimates of variable importance without variable deletion. Figure 5 illustrates the results of the cross-validation of the predicted values and the actual time to failures based on the trained random forest models.

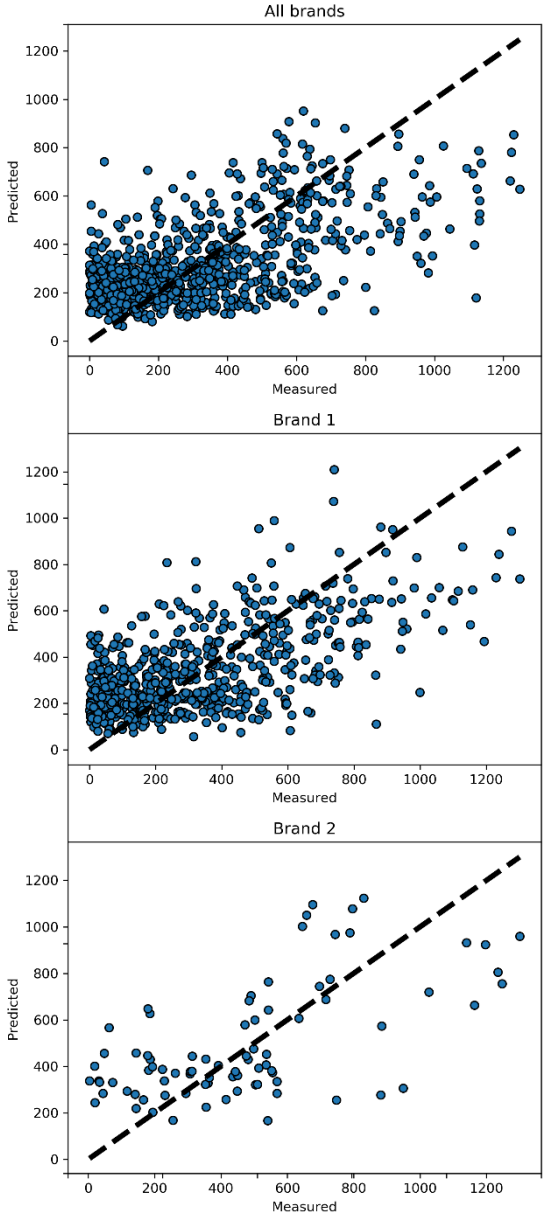

Figure 5 – Time to failure predictions for each model

Figure 6 depicts the performance of the regression models. Figure 6 suggests that developing models for each specific brand would relatively improve the performance of the prediction.

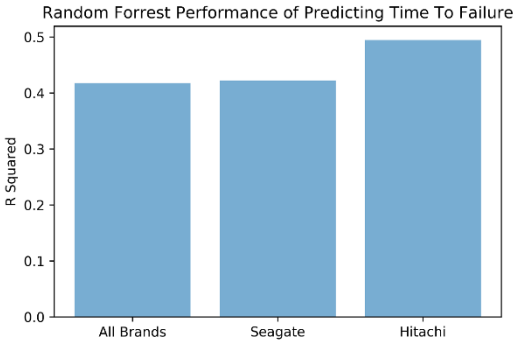

Figure 6 – Performance of regression models

Table 5 ranks the top ten most important variables of each model based on the decrease in mean squared error. The ranking helps to infer the most useful attributes for estimating the time to failure. Except *SMART 9,* which refers to the power-on time of the device and is representative of the devices' age and is quite important in all three models, the rest of the variables do not share the same ranking over the three models. The reason may originate from the fact that different manufacturers may not use the same standards for the same attributes.

Table 5 – Variable importance in each regression model

Variable	Top ten important features ranking		
	All Brands	Brand 1	Brand 2
SMART 9	1	2	1
SMART 240	2	1	
SMART 7	3	4	6
SMART 193	4	6	2
SMART 192	5	8	3
SMART 4	6	10	8
SMART 1	7	5	10
SMART 12	8		9
SMART 242	9	3	
SMART 194	10	9	5
SMART 241		7	
SMART 3			4
SMART 5			7

3.3. Challenges in cloud-based quality monitoring

While the case study presented above illustrates the use of data generated in cloud-based systems for condition monitoring, it also provides the challenges present in doing so. It was previously mentioned that dealing with heterogeneous data types is one of the major challenges of cloud-based manufacturing systems [9]. However, our study suggests that even when utilizing homogenous data types, lack of proper standardization may be a critical hurdle in cloud-based health monitoring. Our segmentation analysis showed that the stats reported by the SMART attributes may correlate differently with each other across various brands. Unified product/services data standards are required for future cloud-based health management systems.

Another important attribute that needs consideration refers to the frequency of data collection. Since data storage and management become challenging for large amounts of data, optimum data collection frequency for each prognostication task should be obtained in order to avoid unnecessary costs related to data computation, data storage, and energy consumption.

4. Conclusions

The emergence of cloud-computing has been pushing toward a paradigm shift in manufacturing, entailing a transition from product-manufacturing plants to service-oriented entities. While the communications with and the usage of the manufacturing resources have been changing toward a Big Data environment, corresponding quality monitoring techniques should adapt accordingly. Since cloud-manufacturing platforms facilitate manufacturing data collection, data-driven condition monitoring techniques could be exploited to seize such opportunities.

It should be noted that the analyses discussed in this manuscript are based upon real-time product health status monitoring via product specific data. However, failures can also occur due to other unpredictable reasons such as consumer misuse, disasters or inappropriate environmental conditions.

This paper provides a case study of a real-world cloud storage health monitoring and failure prediction system. Hard disk drives' SMART attributes have been utilized in order to predict the time to failure of drives in Redundant Array of Inexpensive Disks (RAID) systems of cloud storage centres. Moreover, challenges with respect to data collection frequency and data standardization have been discussed.

Future work should focus on other pattern recognition techniques to improve the performance of predictions. Moreover, data from other cloud-based sectors should be investigated in order to analyze the generalization of the predictions.

Acknowledgements

This work was funded by the National Science Foundation–USA under grant number CBET-1705621. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Xu, X., 2012, "From Cloud Computing to Cloud Manufacturing," Robot. Comput. Integr. Manuf., **28**(1), pp. 75–86.

[2] Wang, X. V., and Wang, L., 2014, "From Cloud Manufacturing to Cloud Remanufacturing: A Cloud-Based Approach for WEEE Recovery," Manuf. Lett., **2**(4), pp. 91–95.

[3] Zhang, L., Luo, Y., Tao, F., Li, B. H., Ren, L., Zhang, X., Guo, H., Cheng, Y., Hu, A., and Liu, Y., 2014, "Cloud Manufacturing: A New Manufacturing Paradigm," Enterp. Inf. Syst., **8**(2), pp. 167–187.

[4] Lee, J., Lapira, E., Bagheri, B., and Kao, H. an, 2013, "Recent Advances and Trends in Predictive Manufacturing Systems in Big Data Environment," Manuf. Lett., **1**(1), pp. 38–41.

[5] Gao, R., Wang, L., Teti, R., Dornfeld, D., Kumara, S., Mori, M., and Helu, M., 2015, "Cloud-Enabled Prognosis for Manufacturing," CIRP Ann. - Manuf. Technol., **64**(2), pp. 749–772.

[6] Lee, J., Ardakani, H. D., Yang, S., and Bagheri, B., 2015, "Industrial Big Data Analytics and Cyber-Physical Systems for Future Maintenance & Service Innovation," *Procedia CIRP*, pp. 3–7.

[7] Wang, L., 2013, "Machine Availability Monitoring and Machining Process Planning towards Cloud Manufacturing," CIRP J. Manuf. Sci. Technol., **6**(4), pp. 263–273.

[8] Mourtzis, D., Vlachou, E., Milas, N., and Xanthopoulos, N., 2016, "A Cloud-Based Approach for Maintenance of Machine Tools and Equipment Based on Shop-Floor

Monitoring," *Procedia CIRP*, pp. 655–660.

[9] Yang, S., Bagheri, B., Kao, H.-A., and Lee, J., 2015, "A Unified Framework and Platform for Designing of Cloud-Based Machine Health Monitoring and Manufacturing Systems," J. Manuf. Sci. Eng., **137**(4), pp. 40914–40916.

[10] Gopalakrishnan, P. K., and Behdad, S., 2017, "Usage of Product Lifecycle Data to Detect Hard Disk Drives Failure Factors," *ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, American Society of Mechanical Engineers, p. V004T05A031-V004T05A031.

[11] Pinheiro, E., Weber, W., and Barroso, L., 2007, "Failure Trends in a Large Disk Drive Population," Proc. 5th USENIX Conf. File Storage Technol. (FAST 2007), (February), pp. 17–29.

[12] Murray, J. F., Hughes, G. F., and Kreutz-Delgado, K., 2005, "Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application," J. Mach. Learn. Res., **6**, pp. 783–816.

[13] Hughes, G. F., Murray, J. F., Kreutz-Delgado, K., and Elkan, C., 2002, "Improved Disk-Drive Failure Warnings," IEEE Trans. Reliab., **51**(3), pp. 350–357.

[14] Zhu, B., Wang, G., Liu, X., Hu, D., Lin, S., and Ma, J., 2013, "Proactive Drive Failure Prediction for Large Scale Storage Systems," *2013 IEEE 29th Symposium on Mass Storage Systems and Technologies (MSST)*, pp. 1–5.

[15] Li, J., Stones, R. J., Wang, G., Li, Z., Liu, X., and Xiao, K., 2016, "Being Accurate Is Not Enough: New Metrics for Disk Failure Prediction," *2016 IEEE 35th Symposium on Reliable Distributed Systems (SRDS)*, pp. 71–80.

[16] Li, J., Stones, R. J., Wang, G., Liu, X., Li, Z., and Xu, M., 2017, "Hard Drive Failure Prediction Using Decision Trees," Reliab. Eng. Syst. Saf., **164**(Supplement C), pp. 55–65.

[17] Pang, S., Jia, Y., Stones, R., Wang, G., and Liu, X., 2016, "A Combined Bayesian Network Method for Predicting Drive Failure Times from SMART Attributes," *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 4850–4856.

[18] Rincón, C. A. C., Pâris, J. F., Vilalta, R., Cheng, A. M. K., and Long, D. D. E., 2017, "Disk Failure Prediction in Heterogeneous Environments," *2017 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, pp. 1–7.

[19] "The Raw Hard Drive Test Data" [Online]. Available: https://www.backblaze.com/b2/hard-drive-test-data.html. [Accessed: 13-Sep-2017].

[20] Maaten, L. Van Der, and Hinton, G., 2008, "Visualizing Data Using T-SNE," J. Mach. Learn. Res. 1, **620**(1), pp. 267–84.

[21] Breiman, L., 2001, "Random Forests," Mach. Learn., **45**(1), pp. 5–32.