Module Code: CSMRS16-22-3A | **Assignment Report Title**: Research project proposal | **Student Number**: 29802635
Date: 2023-01-07 | **Actual hrs spent for the assignment**: 12h

Research Project Proposal

1. Research background and literature review approach

1.1 Research background

As the demand for physical and cloud storage is increasing rapidly, the number of hard disk drives in operation is also increasing and with it so is the number of disk failures. As these failures usually impact the quality of the storage services it is clear that these cannot be ignored and a more proactive approach is required (waiting for a disk to fail before replacing it is more disruptive than replacing the disk before it is about to fail).

Hard Disk Drives (or HDDs) were introduced by IBM in 1956 and since then they have become the most wide-spread technology for data storage. They remain the most popular storage media in data. They remain the most popular storage media in data centers even after the rise of the Solid State Drive (or SSD) – which no longer has moving parts but rather chips with storage cells – because of their price to capacity and life expectancy ratio [2].

With the wide spread adoption of cloud services for workloads ranging from small (ie. individual virtual machines hosting a personal blog) to big (ie. using data science to predict or model weather patterns based on massive data sets collected over decades) it becomes clear that storage systems are required to scale to Petabytes and Exabytes which results in using hundreds of thousands and millions of HDDs per data center. At this scale disk failures are no longer rare events but rather they become the norm and with that comes the need to have optimal strategies to deal with such failures.

It is true that data loss caused by disk failure has been reduced by the adoption of solutions such as redundant arrays of inexpensive disks (RAID) however, when a disk that is part of a storage array fails and is replaced, the recovery process is a lengthy one and while it is running, additional stress is added on the remaining disks which can cause, in the best case scenario, performance degradation of the system, and, in the worst case scenario, data loss caused by the failure of one or more disks in the same storage array. This approach works however due to its reactive nature it remains an unsatisfying solution [2].

In recent years focus has been shifted towards exploring more proactive solutions such as predicting when a HDD is close to failure such that the maintenance window required to replace it can be scheduled in advance to reduce the impact on the overall performance of the system [2, 5].

Due to shifts towards predictive systems, machine learning approaches have been gaining increasing popularity – especially the ones using models trained on S.M.A.R.T. data by relying on internal attributes of HDDs as indicators of drive reliability [2].

1.2 **Literature review approach**

The approach for literature review is around identifying the state-of-the-art approaches to predicting remaining useful life (RUL) for HDDs. Several sources have been identified for the initial literature review each using a different approach for predicting RUL.

Hu et al. [5] proposes a model based on LSTM to predict disk failure in a given interval (30 days before the actual failure). Santo et al. [2] follows recent research in predictive maintenance, provides an overview of State-of-the-Art approaches and presents a deep learning approach to address data sparsity, need for domain knowledge and feature engineering to predict RUL of a HDD by identifying specific health conditions on the basis of S.M.A.R.T. attributes values using three main steps: defining the health degree for each HDD, extracting sequences in a specific time window for each hard disk and then assessing the health status through LSTM by associating a health level to each temporal sequence. The Conf. Paper [3] proposes a fault prediction method based on multi-instance LSTM neural network where the data in the entire degradation process is regarded as a sample then using the LSTM network the time characteristics of the data are mined and finally a multi-instance learning method is used to treat the degradation characteristics of the full-life data as a data bag and divide it into multiple instances thus the entire life cycle data is used for HDD abnormality detection. Coursey et al. [4] proposes methods for data standardization, normalization and RUL prediction using Bidirectional LSTM network with multiple days of look-back period considering S.M.A.R.T. attributes highly correlated to failure and builds a prediction pipeline that takes into consideration the long-term temporal relations in the failure data.

Building on the initial literature review, search engines like Google and DuckDuckGo together with academic journals like IEEE and ScienceDirect will be used to identify more relevant articles in the context of predicting RUL together with a stream based approach at predicting RUL to evaluate methods and tools used by researchers and select the ones that apply to this research proposal.

2. **Research scope**

The proposed research project aims to build a practical application for predicting RUL which will use at least two datasets (Backblaze [1] and at least another one that will be identified later on) for the initial training which will keep itself up-to-date by continuously ingesting new measurements from live/real time S.M.A.R.T. streams of data including from never before seen HDDs with the possible extension to SSDs and NVMEs (provided such datasets can be found freely online).

The first objective is to prepare a diverse enough dataset on which to train the algorithm. The data will be qualitative and sourced from public data sources afterwards it will be processed to handle the outliers by either removing or reweighing their impact by leveraging state-of-the-art techniques.

The second objective (which is also the main one) is to identify the best technique(s) and tool(s) for handling the data and training the model (which will use either LSTM or a mix of algorithms depending on their overall accuracy) for predicting RUL.

Another objective (third) is to measure the efficiency and cost of the chosen Machine Learning algorithm(s) and compare with other state-of-the-art models and techniques.

The last objective (fourth) is the delivery of a practical application that can be used in a production environment to predict RUL with high accuracy in a cost effective manner and with little to no maintenance effort or operational cost.

At the very least, this research project will contribute by reporting the computational and time costs of training and applying the Machine Learning algorithms on this particular type of dataset which will allow repurposing them in the future to other datasets. The main contribution of this project, if successful, will be the practical application.

3. **Intellectual Challenges**

All computer systems require some form of storage (local or remote, SSD or NVME or HDD based, etc) in order to persist data and depending on the importance of the data that is being stored one must have at the minimum one or more form(s) of disaster recovery solution to minimize the risk of data loss (there's a plethora or solutions including but not limited to USB sticks, optical media, tape drives, Network Attached Storage - NAS, cloud backups such as Backblaze[1], etc). Storing multiple copies of the data comes with extra cost and environmental impact (due to more energy and physical resources being needed for the hardware to be made available, shipped, powered on, kept up to date, etc) however this can be reduced considerably if a practical application exists that can predict accurately the remaining useful life of storage media (particularly HDDs, which are still the most cost effective solution, but can be extended to other media such as SSDs, NVMEs, etc in the future) as based on these predictions the owner of the data can proactively migrate the data to new media before a failure occurs rather than needing to store extra copies of the data to cope with an unpredicted failure event.

The proposed research will attempt to establish a model that can predict RUL for HDDs (at first and later on be extended to SSDs and NVMEs) with high accuracy and do so in a practical manner such that the resulting application requires zero to minimal maintenance to operate once deployed to a production environment.

4. **Methodology and research design**

4.1 **Approach and methodology**

The methodology to be adopted is a quantitative empirical experiment which will follow the Extract Clean Transform (ECT) structure:

(a) Extract the data from the data sources

(b) Clean the data, highlight outliers and remove them from the datasets or reweigh them

(c) Create a training and testing dataset (experiment with 60/40, 70/30, 80/20, 90/10)

(d) Train the algorithm(s) on the test data

(e) Evaluate the algorithms using a 10-cross-fold evaluation method and select the most effective one or a combination between them

(f) Compare the selected algorithm(s) with existing results from other papers

(g) Measure computational cost of each step

After the training dataset has been prepared, a number of Machine Learning method(s) will be trained (with LSTM being the main candidate) and compared from a performance and accuracy point of view while at the same time looking at if and how they can handle new data as well as keeping themselves up-to-date training wise while ingesting streams of new data.

The success of the Machine Learning algorithm will be determined by whether it is able to predict RUL with high accuracy (over 90%) on the initial training dataset combined with its ability to maintain high accuracy over time (when predictions start being made by taking into account information that was not used in the initial training but rather information that the algorithm ingested over time and used to train itself) together with the operational cost required for the exercise (compute resources needed and time it takes to train and make predictions, engineering time required for operating the application).

4.2 **Plan of tasks/activities, deliverables and estimated effort**

Tasks	**Deliverables**	**Effort** (person-weeks)
1) Review literature that covers predicting RUL using LSTM with a stream based learning approach - Evaluate the techniques that have been used previously by researchers and identify which	- Design of code to be used to implement the practical application;	- 1 week for the literature review; - < 1 week to write a small POC application to be used as the foundation for the main application;

can be applied to the proposed research		
2) Find more publicly available datasets containing HDDs and SMART measurements - For workload and data diversity at least two datasets need to be used for the proposed research (Backblaze [1] is one of them);	- Create a more generic dataset with more makes and models for disks as well as potentially more workloads;	- 1 week to search for publicly available datasets; - potentially 1 or 2 more weeks in case certain datasets of interest require agreements to be signed;
3) Determine the data analysis tools and techniques to be implemented - Review the state-of-the-art techniques used by researchers evaluating the various use cases and applicability to the proposed research;	- Inform on the state-of-the-art techniques to be used in the creation of the prototype prediction model; - This could facilitate the identification of specific algorithms and tools to be implemented;	- 1 week for the review of the techniques and tools; - 1 week to write descriptions and justifications of the selected tools;
4) Perform data preprocessing - Perform the relevant data pre-processing and data manipulation to develop the required sets (train / test split) for algorithm training;	- The creation of a diverse and balanced dataset with the required variables to be used for predicting RUL;	- 2 – 3 weeks to perform the data preprocessing; - 1 week to write up the processes ;
5) Train the selected algorithm on the training data - Use the data analysis techniques identified in the third task to create the predictive model;	- A model capable of predicting RUL of HDDs for which S.M.A.R.T. measurements have not been seen by the algorithm during training;	- 1 - 2 weeks to train the model;
6) Evaluate the performance of the created model - Implement an evaluation method, such as k-fold cross validation, to determine if the goal accuracy can be achieved;	- Evaluation metrics such as average accuracy over several iterations; - This allows conclusions to be drawn on the performance of the model;	- 1 week to evaluate the performance; - 1 week to write the findings;
7) Discuss results and findings - Critique of the results produced and potential strengths/weaknesses of the overall project; - Determine areas for related	- Insights into results, findings, highlighting areas of interest;	- 2 weeks to write the discussion of results;

future work;		
Total		13 - 17 weeks
Resources required by the project:	Sufficient computer processing power to develop the predictive model and sufficient storage capacity for the data that needs to be processed e.g., through algorithm training and evaluation together with the S.M.A.R.T. datasets	
Costing for this project (if any):	No major costs as the required hardware is already available and the datasets that will be used are free.	

5. **Ethical and risk considerations**

The dataset that sits at the foundation of this project (Backblaze [1] together with any additional datasets that will later on be used) contains a list of hard disk drive models together with measurements at different points in time for their S.M.A.R.T attributes, is completely anonymous by nature (hardware telemetry) and does not contain any personal identifiable information. Given that the dataset contains entries for a limited number of hard disk makes and models as well as the fact that it is imbalanced (there are considerably more entries for certain makes and models compared to others) there is the potential for bias however this will be accounted for by training the model individually per make and model using the data that is considered valid as well as training the model using the complete dataset stripped of any make and model information such that when attempting to predict remaining useful life of a disk of a particular make and model which has not been seen by the model a general prediction can be made.

To conclude, there are minimal potential issues concerning ethics with the dataset due to it being publicly available, there are no ethical practice issues as we're relying on data that has already been collected and there are no legal issues as this data is at its core telemetry for hardware that has been made public by its owner for research purposes.

6. **Conclusion and Reflection**

Overall, this project proposal aims to use existing methodology and tools to build a practical application for predicting RUL of HDDs using LSTM and/or a mix of other Machine Learning algorithms in such a way that the application is able to make predictions with high accuracy, keep itself up-to-date and have a low overall operational cost.

One issue that this research proposal has to deal with is in the context of the dataset which, based on the initial research, is expected to be highly imbalanced (all datasets containing S.M.A.R.T. measurements will contain considerably more data about disks in good working order compared to

disks that are about to fail or have failed already not to mention that datasets are limited to the disk models used by the entity that generated the dataset so no one dataset will contain information about each available HDD model).

Another issue with the proposal is that it aims to compare a number of Machine Learning techniques including creating combinations with the aim of identifying one or more which can accurately predict RUL of HDDs which is expected to be time consuming.

The final and possibly the biggest issue is that the dataset needed to train the model(s) comes from one entity (Backblaze [1]) which uses the HDDs for a specific type of workload and has a specific process for identifying and replacing the failed drives not to mention it uses a limited set of disk models. The success of this research proposal depends on finding at least one more dataset that can be merged with the existing one and used for training such that the model can be trained on data that is more generic with respect to the workload and disk model.

References

[1] A. Klein, "Backblaze Drive Stats for Q2 2022", Aug. 2019, [online] Available:https://www.backblaze.com/blog/backblaze-drive-stats-for-q2-2022/.

[2] A. De Santo, A. Galli, M. Gravina, V. Moscato and G. Sperlì, "*Deep Learning for HDD Health Assessment: An Application Based on LSTM*", in IEEE Transactions on Computers, vol. 71, no. 1,pp. 69-80, 1 Jan. 2022, doi: 10.1109/TC.2020.3042053.

[3] "*A multi-instance LSTM network for failure detection of hard disk drives*", 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), 2020, pp. 709-712, doi:10.1109/INDIN45582.2020.9442240.

[4] A. Coursey, G. Nath, S. Prabhu and S. Sengupta, "*Remaining Useful Life Estimation of Hard Disk Drives using Bidirectional LSTM Networks*", 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 4832-4841, doi: 10.1109/BigData52589.2021.9671605.

[5] Lihan Hu, Lixin Han, Zhenyuan Xu, Tianming Jiang and Huijun Qi, "*A disk failure prediction method based on LSTM network due to its individual specificity*", Procedia Computer Science, vol. 176, pp. 791-799, 2020.