Exploratory Data Analysis for Data Center Energy Management

Wania Khan
Leeds Beckett University
Leeds, UK
w.khan2776@student.leedsbeckett.ac.uk

Davide De Chiara
ENEA-ICT Division
R.C. Portici, Italy
davide.dechiara@enea.it

Ah Lian Kor
Leeds Beckett
University Leeds, UK
a.kor@leedsbeckett.ac.uk

Marta Chinnici
ENEA-ICT Division
R.C. Casaccia, Italy
marta.chinnici@enea.it

ABSTRACT

The continuous improvement in energy efficiency of existing data centers would help reduce their environmental footprints. Greening of Data Centers could be attained using renewable energy sources or more energy efficient compute systems and effective cooling systems. A reliable cooling system is necessary to generate a persistent flow of cold air to cool servers that are subjected to increasing computational load demand. As a matter of fact, servers' dissipated heat effects a strain on the cooling systems and consequently, on electricity consumption. Generated heat in the data center is categorized into different granularity levels namely: server level, rack level, room level, and data center level. Several datasets are collected at ENEA Portici Data Center from CRESCO 6 cluster – a High-Performance Computing Cluster. The cooling and environmental aspects of the data center is also considered for data analysis. This research aims to conduct a rigorous exploratory data analysis on each dataset separately and collectively followed in various stages. This work presents descriptive and inferential analyses for feature selection and extraction process. Furthermore, a supervised Machine learning modelling and correlation estimation is performed on all the datasets to abstract relevant features. that would have an impact on energy efficiency in data centers.

Keywords

Data Center; HPC; Data Mining; Big Data; Thermal; Hotspot; Cooling; Thermal Management; Descriptive Analysis; Inferential Statistics; Linear Regression Modelling; Data Analysis.

1. INTRODUCTION

A huge proportion of worldwide generated electricity is produced by hydrocarbon combustion which causes a rise in carbon emission and GHGs. Data Centers (DC) worldwide were estimated to have consumed between 203 to 271 billion kWh of electricity in the year 2010 [1]. Energy use continues that is, increasing by 4% from 2014-2020. Based on current trend estimates, U.S. data centers are projected to consume approximately 73 billion kWh in 2020 [2]. Over 90% of the energy input is being dissipated as waste heat energy and also due to energy consumption by cooling systems. Thus, both cooling and compute systems, have been critical targets for energy savings. Thermal mismanagement in a DC could be the primary contributor to thermal degradation of compute systems. CPUs are the primary energy consumers and waste heat dissipators [3] and often, it is necessary to disperse dissipated waste heat so that there will be an even distribution of waste heat within a premise to avoid heat islands. An approach to avoid overheating are thermal-aware schedulers with system-level work placements [4] as for example execute 'hot' jobs on 'cold' compute nodes; adopt predictive model for job schedule selection [5]; adopt ranked node queue based on thermal characteristics of rack layouts and optimization. Heat modelling provides a link for server energy consumption and their associated waste heat. Thermal-aware monitoring acts as a thermal-eye for the scheduling process and entails recording and evaluation of heat distribution within DCs [6]. Thermal profiling is based on useful monitoring information on workload-related heat emission and is useful to predict the DC heat distribution.

The aims and objectives of this research is to perform an exploratory data analysis on the datasets collected for an entire year of 2020 from the ENEA Portici data center, Italy. This research involves abstraction of most relevant data features which could be fed further advanced data analytics for data center energy consumption and resources utilization. The following objectives are formulated to achieve the aim of this research:

1. **Research scoping**: this study mainly focuses on preliminary stages of data analysis. Also, the dataset is collected from one specific cluster i.e., CRESCO6 at ENEA portici data center.
2. **Data Collection:** data collected from four different sources of the data center with the data dictionary to develop a better understanding about dataset description.
3. **Data Preprocessing and Appreciation:** implement data cleansing process on each dataset followed by exploratory data

analysis. Descriptive and inferential statistics are implemented to extract relevant features.

4. **Data Modelling:** build machine learning models to explore the relationships between extracted features in (3) and their changes with respect to time.

5. **Results Analysis:** interpretation of results and final discussion on (4).

In this paper, our analysis explores the relationship between thermal data (incl. environmental data and cooling machine data) and computer workload data. The novel contribution of the research is the use of actual and real big dataset for ENEA High Performance Computing (HPC) CRESCO6 compute nodes. A supervised learning technique has been employed to explore the relationship between different attributes of different datasets. The aim of this research is to improve the thermal aspect as well as energy consumption of the datacenter. The paper is organised as follows: Section I – Introduction; Section II – Background: Related Work; Section III – Methodology; Section IV – Results and Discussion; Section V – Conclusion and Future Work.

2. RELATED WORK

DC energy efficiency has been a long-standing challenge due to many factors that affect DC energy efficiency and it is the trade-off between performance in the form of productivity and energy saving. Interesting trade-offs between geolocations and DC energy input requirements (e.g., cold geolocations and free air-cooling; hot, sunny geolocations and solar powered renewable energy) are yet to be critically analysed [7]. One of the thermal equipment-related challenges is raising the setpoint of cooling equipment or lowering the speed of CRAC (Computer Room Air Conditioning) fans to save energy. Another long-standing challenge is IT resource over-provisioning that causes energy waste due to idle servers [22]. Significant research investigates optimal allocation of PDUs (Power Distribution Units) for servers, multi-step algorithms for power monitoring, and on-demand provisioning reviewed in [7]. Other related work addresses workload management, network-level issues as optimal routing, Virtual Machines (VM) allocation, and balance between power savings and network QoS (Quality of Service) parameters as well as apposite metrics for DC energy efficiency evaluation. One standard metric used by a majority of industrial DCs is Power Usage Effectiveness (PUE) proposed by Green Grid Consortium [8]. It shows the ratio of total DC energy utilisation with respect to the energy spent exclusively by IT equipment. A multitude of efficiency metrics evaluate the following:

- thermal characteristics;
- ratio of renewable energy use;
- energy productivity of various IT system components, and etc.

There is a pressing need to provide a holistic framework that would thoroughly characterise DCs based on a fixed set of metrics and reveal potential pitfalls in their operations. Though some existing research work has made such attempt but to date, we are yet to have a standardised framework [9, 10]. To reiterate, the thermal characteristics of the IT system ought to be the primary focus of an energy efficiency framework because it is the main energy consumer within a DC. Several research have been conducted to address this issue. Sungkap et al. [11] propose an ambient temperature-aware capping to maximize power efficiency while minimising overheating. Their research includes an analysis of the composition of energy consumed by a cloud-based DC. Their findings for the composition of DC energy consumption are approximately 45% for computer systems; 40% for refrigeration-based air conditioning; remaining 15% for storage and power distribution systems. This implies that approximately half of the DC energy is consumed by non-computing devices. In [12], Wang and colleagues present an analytical model that describes DC resources with heat transfer properties and workloads with thermal features. Thermal modelling and temperature estimation based on thermal sensors data ought to consider the emergence of server hotspots and thermal solicitation due to the increase in inlet air temperature, inappropriate positioning of a rack or even inadequate room ventilation. Such phenomena are unraveled by thermal-aware location analysis. The thermal-aware server provisioning approach is presented in [13] to minimise the total DC energy consumption which considers the maximum allowable working temperature of the servers. Typical thermal-aware scheduling algorithms are reactive, proactive or mixed. However, there is no reference to heat-modelling or thermal-monitoring and profiling. Kong [3] highlights important concepts of thermal-aware profiling, thermal-aware monitoring, and thermal-aware scheduling. Thermal-aware techniques are linked to the minimisation of waste heat production, heat convection around server cores, task migrations, and thermal-gradient across the microprocessor chip, and microprocessor power consumption. Dynamic thermal management (DTM) techniques in microprocessors encompasses the following: Dynamic Voltage and Frequency Scaling (DVFS), Clock gating, task migration, and Operating System (OS) based DTM and scheduling. In [14], Parolini proposes a heat model; provides a brief overview of power and thermal efficiency from microprocessor micro-level to DC macro-level. To reiterate, it is essential for DC energy efficiency to address thermal awareness in order to better understand the relationship between both the thermal and the IT aspects of workload management. In this paper, the authors have presented how to select and extract features from thermal, environmental and compute nodes sensor datasets of a HPC data center cluster. Subsequently, a link to the DC energy consumption is made. This research involves measurement and analysis of compute nodes sensor and refrigerating machines. Overall, an effective DC management requires energy use monitoring, particularly, energy input, IT energy consumption, monitoring of supply air temperature and humidity at room level (i.e., granularity level 0 in the context of this research), monitoring of air temperature at a higher granularity level (i.e., at Computer Room Air Conditioning/Computer Room Air Handler (CRAC/CRAH) unit level, granularity level; **1)** Measurements taken are further analysed to reveal the extent of energy use and economisation opportunities for the improvement of DC energy efficiency level (granularity

level); **2)** DC energy efficiency metrics will not be discussed in this paper because they have been discussed in [21]. However, the discussion in the subsequent section primarily focuses on thermal guidelines from American Society of Heating, Refrigerating and AC Engineers (ASHRAE) [15].

3. METHODOLOGY

To determine the energy consumption in the data center, first it is important to critically analyse the data at each granularity level (which has been discussed in previous section). This research is conducted in collaboration with ENEA Portici Data Center, Italy and it is specifically focused on the data collected from CRESCO6 cluster. In this work, an exploratory data analysis is performed on different datasets individually and collectively as well. The proposed methodology aims to address different stages in a data lifecycle – data cleansing, data aggregation, data synchronisation, data transformation into useful information for further and advanced data modelling. Our research methodology is categorized into macro-methodology and micro-methodology levels.

3.1 Macro-Methodology

Macro-methodology (fig. 1) depicts an overview of the stages that are involved for our data analysis. Three methodologies are taken into consideration to explore which one will be appropriate for this study:

1. KDD – Knowledge discovery in databases Process: a high-level interactive and iterative process of using data mining method for non-trivial extraction of implicit and useful information from databases. [16]
2. CRISP-DM – Cross Industry Standard Process for Data Mining: a process developed for forming an industry standard. [19]
3. SEMMA - Sample, Explore, Modify, Model and Assess: data mining process of highly iterative nature developed by SAS Institute. [20]

After a comprehensive analysis of all three methodologies and comparison performed in study [16], it is clear the SEMMA process consists of practical implementation of all the five stages as KDD process and both of the processes provides a high-level overview of the data lifecycle process. The data processing steps of SEMMA and KDD are somehow similar; However, SEMMA provides data modelling steps to reliably predict the final, desired outcome of the process. Furthermore, SEMMA does not consider the business aspect of the study unlike CRISP-DM. Since, this research is mainly aimed to perform the exploratory data analysis on the compute data, thermal data and environment data of ENEA Portici data center to determine relevant data features for energy management. Hence SEMMA is implemented as a macro level methodology for this study which will be further analyzed in-depth at each stage in the micro-methodology section.

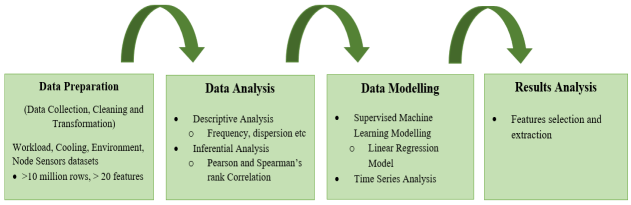

Fig. 1. Methodology Phases.

The proposed methodology is divided into four major phases which overlaps with a typical data lifecycle. The following section provides an overview of the stages which will be discussed in detail in the later sections:

a. Data Preparation
   1. Data Cleaning: involves deletion of null rows, columns, and error values from the datasets.
   2. Data Transformation: sampling and aggregation of datasets.

b. Data Analysis: The data analysis phase will encompass 3 levels of data analysis namely:

1. Descriptive Statistical Analysis: describes about the dataset including min, max and standard deviation of the numerical features.
2. Inferential Statistical Analysis: explores correlation among different attributes of datasets.
3. Machine Learning: linear regression modelling and time series analysis.

3.2 Micro-Methodology

The micro methodology allows in-depth implementation and analysis of each stage defined in macro-methodology such as data cleaning, data collection process and exploratory data analysis etc. The first starting step for this study process involves data dictionary followed by collection process.

3.2.1 Data Dictionary

The HPC Cluster – CRESCO6 initially has about 220 calculation nodes which is subsequently increased to 440 calculation nodes during the period 2019 to 2020. CRESCO 6 is a very high-performance system and consists of 434 calculation nodes for a total of 20832 cores. Each calculation node has 2 intel Xeon platinum 8160 each with 24 cores and 192gb of ram, so 4gb for each core. The nodes are interconnected by an Intel Omni-Path network with 21 Intel Edge. The consumption of electrical power during massive computing workloads is approximately to 190 kW. Following table 1 give information about all four different datasets collected at the data center.

Table. 1. Data Dictionary of all Datasets.

Workload Dataset Attributes	Compute Node Sensors Dataset Attributes	Environment Dataset Attributes	Cooling Dataset Attributes
Jobid = ID of LSF job **Numcores** = number of used cores **User**=name of user **Queue**= name of the queue for job running **Directory**= executable file directory **Executable**= name of the executable file **Job status**= final status of the job **Start**= start of the job execution **Stop**=stop of the job execution **Numhost**= number of hosts engaged	**Time:** record time of readings **Sys_power**: Total instantaneous Power of the node **Resource Consumption Power**: CPU power + Memory Power of the node **Average Fan Speed**: avg speed of the 10 cooling fans installed in the node in RPM (revs per minute) **Sys_util**: percentage of use of the system for single node **Resource Utilisation:** CPU utilisation + Memory utilisation of the node **Average CPU_Temp**: CPU temperature in °C **Average Node_Temp**: Node temperature in °C **Sysairflow**: air flow of the node measured in CFM - indicates the flow of air moved **DCenergy**: energy meter consumed up to the time of reading	**Time:** record time of readings **Average Hot-Aisle Temperature**: temperature in °C **Average Hot-Aisle Humidity**: humidity in percentage **Average Cold-Aisle Temperature**: temperature in °C **Average Cold-Aisle Humidity**: humidity in percentage	**Time:** record time of readings **Machine Name** **Machine status:** working status, on or off **Supply air** **Return Air** **Relative Humidity** **Fan speed of cooling machine** **Cooling**

3.2.2 Data Collection

The first starting step for this study process involves data collection. The data is recorded from CRESCO6 HPC cluster in ENEA Portici datacenter from four different areas i.e., workload data, cooling data, environmental data, and compute nodes sensor data through various internal and external sensors for a period of an entire year from January 1st,2020 till December 31st, 2020.

3.2.3 Data Collection Process

1. Workload Data: This data is collected from the running HPC (High Performance Computing) nodes in CRESCO6 cluster (ENEA Portici DC) whenever it receives a job request from the user. The data contains information of individual jobs submitted by users who use the Cresco6 cluster which is obtained from the job scheduler LSF (Load Sharing Facility) – a workload management platform, job scheduler, for distributed high-performance computing.
2. Environmental Data: The environmental parameters are measured through external sensors for collection of humidity and temperature data inside the room where all the HPC nodes are running. The sensors are installed on both the cold side isle and on the hot side isle near the cluster.
3. Cooling Data: There are three air conditioners operating in CRESCO6 to cool down the HPC nodes running in the ENEA Portici data center. The temperature of the surrounding air is recorded i.e., cold air produced by the conditioning machines and the warm air coming out of the operating nodes. It also consists of data related to humidity and the working intensity of the air conditioner.
4. Compute Nodes Sensor Data: There are several sensors installed in every single HPC node of CRESCO6 cluster through which all the data has been collected.

3.2.4 Data Exploratory Analysis

All the above datasets are applied to data analysis techniques for exploration process. The following sub-sections discusses data preprocessing of each dataset.

3.2.4.1 Data Preparation

For the data cleaning process, redundant and erroneous data is identified in all the datasets. On observing the availability of missing values, we find a consistency in the missing pattern to see if the data can be imputed. Since the missing pattern showed a consistency, so it cannot be imputed. So, all those null values are not considered for further processing and hence removed from the datasets.

3.2.4.2 Data Transformation

All the data values are recorded for each second. To facilitate easier analysis, the data is aggregated into hours via a transformation process that involves the calculation of the mean for each data aggregate. The reason to choose mean value over median is that there is no clear outlier in the dataset and distribution of data is symmetrical [17].

3.2.4.3 Data Descriptive Analysis

Starting with the first dataset, Workload Data, it comprises compute nodes sensor data, Platform Load Sharing Facility (or simply LSF) - a workload management platform, job scheduler, for distributed high-performance computing. A sample of workload dataset is first analyzed to examine its features along with their corresponding values. It covers attributes such as id, job id of LSF job, number of cores, directory, start and stop time of a job executed by a single node, the execution status etc. The total number of features in the dataset are 12 and the total count of values recorded on seconds basis for a year is 4656109.

As for the Environmental data, a sample of environment dataset is first analyzed to examine its features along with their values. This dataset comprises several features which include timestamps, temperature of hot and cold aisles as well as humidity of hot and cold aisles. The total number of features in the dataset is 22 and total count of values is 35579. For numerical data samples in the dataset, central tendency and variations using mean, median and standard deviation for all the features are determined using descriptive data analysis. As all the data entities of environment dataset are equal in number and are not null which indicates the absence of missing values in the dataset. Second-based data are aggregated into hourly-based data. All the hot-aisle (and cold-aisle) temperature and humidity data features are merged into one temperature and humidity feature.

Regarding the Cooling Data, a sample of this dataset is also analyzed separately to examine its features along with their values. The dataset comprises 9 features which includes AC machine name, machine status, in and out supply of air and surrounding humidity and total count of their values is 310245 etc. There is a difference observed between the length of attributes. It is due to the presence of null values. We first determine the null data values in

all the features. To resolve this issue, all the missing rows of the dataset are not considered for further data processing and analysis. Once again, the second-based data is replaced with aggregated hour-based data. Regarding the Compute Nodes Sensor data, a sample of dataset is analyzed to examine its features and their values. The dataset comprises several features which includes timestamp, CPU power and memory power, CPU utilization, fan speed etc. This data consists of 26 features with a total of 11473368 values. Since sensor data is collected on a monthly basis, so all the recorded data values for each month are different. Thus, all the datasets are merged into one dataset for easy handling. The second-based data is aggregated to hour-based data. All the numerical attributes of the collected dataset have object data type which complicates the preprocessing of the dataset. Firstly, all the numerical attributes with non-numeric data type are identified for each month from January 2020 till December 2020 and data type conversion process is employed to convert the object to float attribute. After conversion, it is saved as a new data frame for further analysis.

3.2.4.4 Inferential Data Analysis

Inferential analysis is performed on all the datasets to determine the correlation amongst the features of same dataset as well as between different datasets. Pearson's correlation, and spearman's correlation methods are used in the following sections. In this research, there are four datasets which consists of both normally or non-normally distributed continuous data. In order to perform correlation analysis on both types of datasets, these two correlation coefficients are chosen because Pearson correlation works better with normally distributed data while continuous data with non-normally distribution or outliers can be handled by Spearman rank correlation [18]. The criteria to choose important features is based on the value of correlation coefficients. If this coefficient shows a high positive value, then that specific feature will be selected else, it will not be considered for further analysis.

- **Workload Dataset**

As the workload dataset comprises data about jobs submitted by users which includes both categorical and numerical features. So, the correlation analysis is performed only on numerical data features. The correlation among the workload data features is not good except few attributes shown in fig 2. Both the axis of correlation matrix represents data features. All the attributes will be selected based on their high correlation for further analysis. Now, the correlation of workload data is analyzed with other dataset attributes. Starting with the:

- *Correlation with Environment Dataset:* as this dataset is collected from the sensors placed in the room where servers are running. On applying the correlation between temperature and humidity data with the job execution time, it gives the lowest correlation values as shown below. Since both the data sources are not directly dependent on each other so we can expect the poor relationship amongst them. The reason to perform Pearson's and Spearman's correlation analysis together is to analyse the significant difference between both results. As per the fig 3. we can see that there is not much difference in the results. So, we can use either of them for further analysis.

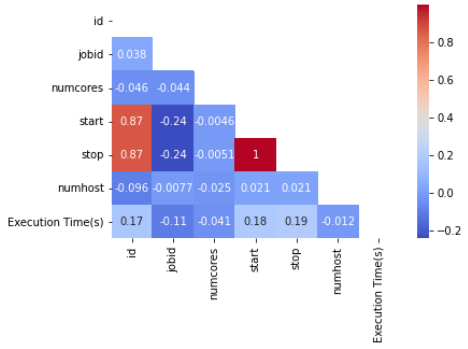

Fig. 2. Correlation within Workload dataset features.

```
    Correlation between Workload Data and Environment Data

Pearsons correlation between Hot-Aisle Temperature and Job execution time: -0.126
Spearmans correlation between Hot-Aisle Temperature and Job execution time: -0.070

Pearsons correlation between cold-Aisle Temperature and Job execution time: -0.100
Spearmans correlation between cold-Aisle Temperature and Job execution time: -0.183

Pearsons correlation between Env Humidity and Job execution time: -0.101
Spearmans correlation between Env Humidity and Job execution time: -0.130
```

Fig. 3. Correlation between Workload dataset and Environment dataset.

- *Correlation with Cooling Dataset:* the estimation of correlation between workload data and cooling machine data result in poor correlation between both dataset features. Since compute nodes are responsible for processing workload while cooling machine provides cooling to the room to maintain the room temperature. Hence, it shows no direct relation but indirectly somewhat.
- *Correlation with Compute Nodes Sensor Dataset:* The estimation of correlation between job execution time and compute nodes sensor data features results in negative values. It gives inverse relationship between the features. Therefor it is not included here.

- **Environment Dataset**

As all the external sensors are placed in the same room at different locations. So, the change in room temperature and humidity will have an approximately same impact on all the sensors. The correlation between the temperature of the aisles is quite high. As the temperature in hot and cold aisle changes together based on the room conditions. The higher the temperature in hot aisle, the colder air will be directed to the front of servers. On the other hand, temperature and humidity are highly negative correlated irrespective of their aisles. This inferential analysis

helped in determining the correlation of these features with the job dataset in the previous section.

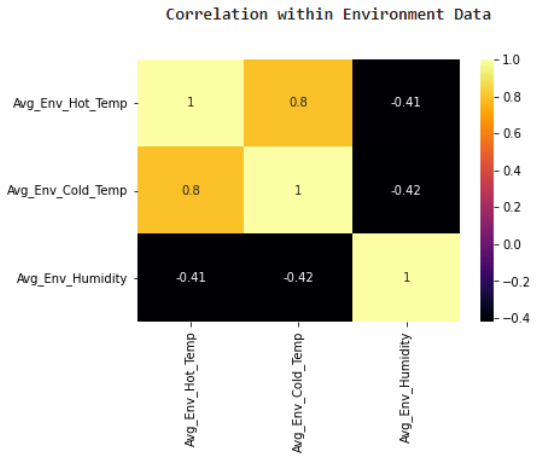

Fig. 4. Correlation within Environment data features.

The correlation analysis is performed between environment data and other datasets as well.

- *Correlation with Workload Data*: we have already discussed in the previous section.
- *Correlation with Compute Nodes Sensor Data:*

The compute nodes sensor data is showing a good correlation with the environment data for most of its features.

The reason behind this is that operation environment for server nodes directly depends on the room temperature and humidity. So, whenever the CPU temperature of the nodes rises due to heavy computations, the environment temperature also increases. Hence, more airflow is provided to the hot aisle of the rack to normalize the operating conditions of room for servers.

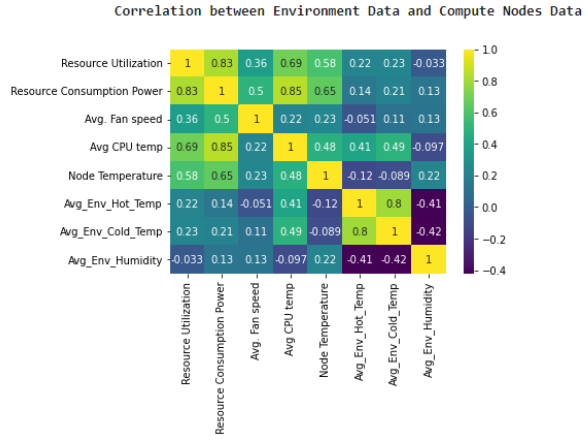

Fig. 5. Correlation between Environment dataset and Compute Nodes Sensor dataset.

- Correlation with Cooling Data:

On observing correlation analysis between environment data features and cooling data features, only one attribute shows a positive correlation with most of the cooling data attributes. While the environment temperature showed a negative correlation with cooling data attributes which presents an inverse relation between them. As it is obvious that whenever the temperature of the room rises due to heavy load computation, the temperature of the air cooling produced by cooling machine will be low.

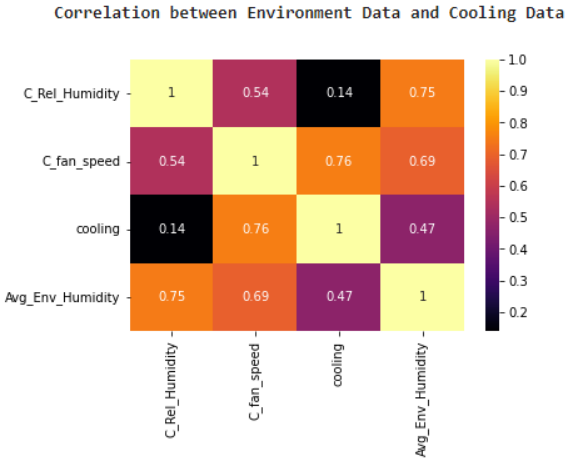

Fig. 6. Correlation between Environment dataset and Cooling dataset.

- **Cooling Dataset**

The correlation between all the features is estimated using Pearson Correlation method. As it is clear from the plot attached, that only few features of cooling dataset i.e., fan_speed and cooling are highly correlated with each other with the highest correlation coefficient of +0.75. While the other attributes showed poor correlation and hence not included in the matrix. This analysis has provided the most relevant features of the cooling dataset that will be used in further data processing stages.

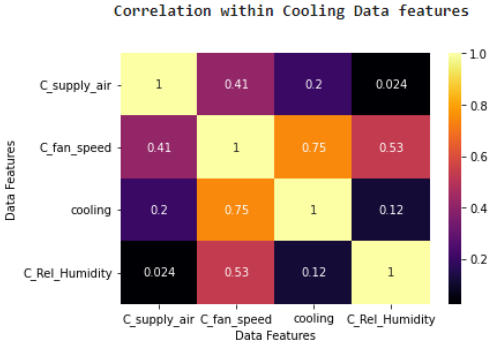

Fig. 7. Correlation within Cooling dataset.

- Correlation analysis of cooling data with workload dataset and environment dataset is already performed in the previous sections.

- Correlation with Compute Nodes Sensor Data: on observing the correlation between cooling data and compute nodes sensor data, only one feature of compute nodes sensor data has shown a slight correlation with the cooling features i.e, fan speed of the server. The reason behind this is that computational nodes in data center are not directly linked with the cooling machine but somehow indirectly they are connected. When the server temperature rises, the speed of the server fan gets faster and in the meanwhile cooling machine also provides more cold air to cool down the hot regions in room.

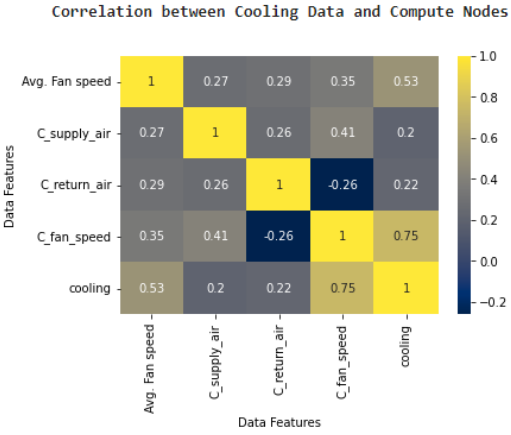

Fig. 8. Correlation between Cooling data and Compute Nodes Sensors data.

- **Compute Nodes Sensor Dataset**

To identify the most relevant features of compute nodes sensor dataset, correlation analysis between all the features of sensor's dataset is performed. And only those features are considered for further processing that have high correlation among them.

- Most of the compute dataset features has shown a good correlation with each other see fig 8. It is apparent that high resource utilization will ultimately consume more power for processing and the CPU temperature will also increase with high and long-term computation. The correlation of compute nodes sensor dataset with other datasets are given in the previous section.

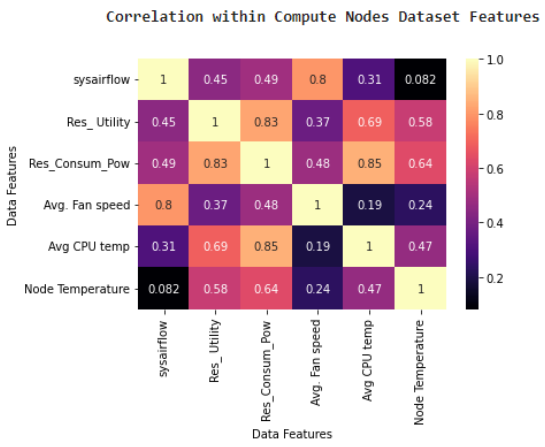

Fig. 9. Correlation within Compute Nodes Sensor Dataset Features.

4 DISCUSSION&RESULTS

This section summarizes all the results obtained through the analysis and provides a brief overview on the preparation of selected data features for the subsequent phase of this research. The main aim of this study is to perform the exploratory data analysis on a collected dataset for feature extraction and selection process. To perform an overall analysis of system operational data with other data attributes, a scatter plot is used as data visualization tool which allows visual analysis of all the data points across a regressed diagonal line. For the selection of data features, a supervised Machine Learning model - linear regression model is implemented on each dataset individually as well as with other datasets. As we have already measured the correlation between the data features in the previous section, the reason to perform a linear regression is to determine the linear relationship between set of dependent and independent variables of the datasets. In this case, all the cooling and environmental parameters are somehow dependent on the computational node. The resources utilization directly depends on the power consumption for the computation. Long-term computation can cause overheating of the server and brings about a rise in the temperatures of the rack and room. The linear regression modelling enabled us to estimate the effect that one changing independent variable has on the dependent variable. First the linear regression modelling is performed for the compute nodes sensor dataset features including resources utilization, resource consumption power and CPU temperature etc. Next their interdependencies are determined by implementing correlation analysis with other datasets including environment data and cooling data. From the following scatter plots in fig. 10, we have analyzed that all the data points near to the diagonal line shows the higher R square value i.e., r2_score of the regression model while the higher dispersion of data points (away from the line) in the graph represents weaker goodness fit of the model. Therefore, based on

the results analysis, only those attributes are selected from all four datasets which showed a linear relationship with least dispersion of data points in the plots and positive correlation with each other as shown in the fig. 10.

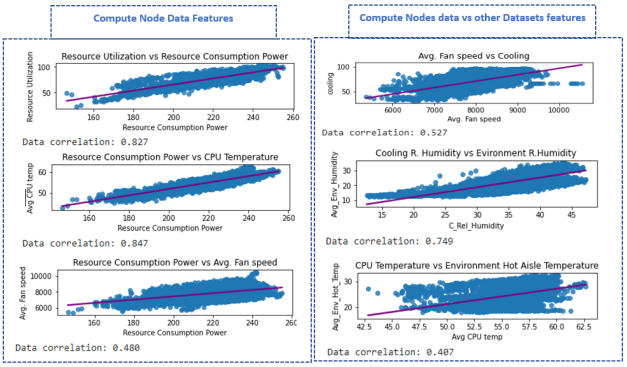

Fig. 10. Linear Regression Modelling and Correlation Analysis between Compute Nodes Sensor Data and other Datasets.

- Time Series Analysis

The most important feature of all in compute nodes sensor dataset are resource utilization, resource energy consumption, average CPU temperature and fan speed. To analyze the behavior of running nodes within its surrounding, all the relevant features including resources utilization, resources power consumption and surrounding temperature are applied to time series analysis for each month. This helped in understanding the characteristics and behavior of the servers in cluster node.

Fig. 1. Time Series Analysis of Compute Nodes Sensor Dataset features.

From the graph plot (fig. 11), it is clear that the resources utilization and power consumption behavior of compute nodes show a seasonality over the given time period. For better understanding of underlying pattens, the time series data is decomposed into several components. The results from time series decomposition in fig 12 are presented for compute nodes sensor data for the period of two months i.e., January and August 2020 below where it can be seen that the trend and seasonality information extracted from the series does seem reasonable. The residuals are also interesting, showing periods of high variability over the time of the series. Furthermore, Augmented Dickey-Fuller (ADF) test is performed on the resource utilization feature which showed that the given time series is stationary. These results helped us in disentangle the time series to make it easier to understand. The exploratory data analysis and time series decomposition implemented in this research enables future forecasting of energy consumption and efficient thermal management in data centers.

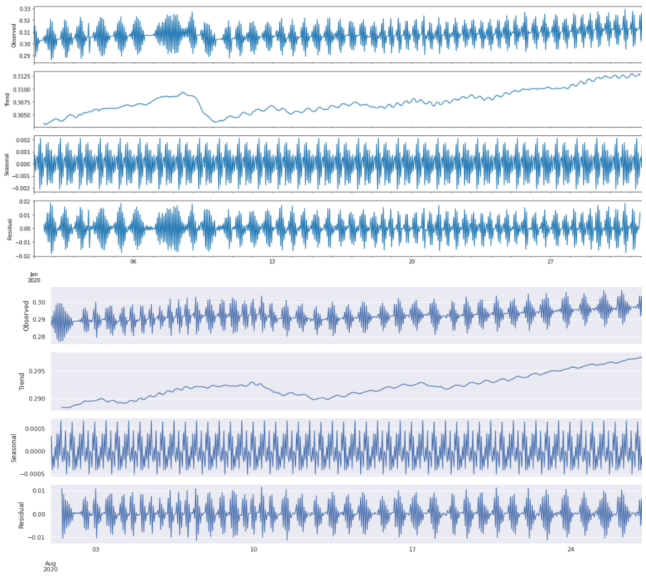

Fig. 12. Time Series Decomposition of Resources Utilization in DC.

5. CONCLUSION AND FUTURE WORK

In order to support sustainable development goals, energy efficiency ought to be the ultimate goal for a DC with a sizeable high-performance computing facility. To reiterate, this work primarily focuses on two major aspects: IT equipment energy productivity and thermal characteristics of an IT room and its infrastructure. The findings of this research are based on the analysis of available monitored thermal characteristics-related data for ENEA-HPC DC, CRESCO6. These findings will be used for advanced data analytics modelling of energy consumption and resource utilization in data center for efficient energy management in future. In this research, preprocessing of unstructured datasets of CRESCO6 IT room is performed through exploratory data analysis including data cleansing, descriptive and inferential statistical analysis. Additionally, a supervised Machine Learning model i.e., linear regression modelling is performed on four different datasets individually and collectively to determine the linear relationship between dependent and independent data features. Furthermore, correlation matrix and scores are used to observe the strong positive correlation between the data features. The criterion we have used to perform feature extraction is based on the fact that correlation must show a positive and linear relationship with the dependent data features. Overall, this research has discussed an

implementation of feature selection and extraction process through various exploratory data processing stages in order to choose the most relevant data attributes for efficient energy management in data center.

REFERENCES

[1] Jonathan G Koomey. 2008. Worldwide electricity used in data centers. Environmental Research Letters 3, 3 (2008), 034008. DOI: https://doi.org/10.1088/1748 9326/3/3/034008

[2] United States Data Center Energy Usage Report | Energy. Retrieved May 28, 2022 from https://eta.lbl.gov/publications/united-states-data-center-energy

[3] Joonho K., Sung Woo C., and Kevin S.. 2012. Recent thermal management techniques for microprocessors. ACM Comput. Surv. 44, 3, Article 13 (June 2012), 42 pages. https://doi.org/10.1145/2187671.2187675.

[4] Tobias Van Damme, Claudio De Persis, and Pietro Tesi. 2019. Optimized Thermal-Aware Job Scheduling and Control of Data Centers. IEEE Transactions on Control Systems Technology 27, 2 (2019), 760-771. DOI: https://doi.org/10.1109/tcst.2017.2783366

[5] Georgios Varsamopoulos, Ayan Banerjee and Sandeep K. S. Gupta. 2009. Energy Efficiency of Thermal-Aware Job Scheduling Algorithms under Various Cooling Models. Communications in Computer and Information Science (2009), 568-580. DOI: https://doi.org/10.1007/978-3-642-03547-0_54

[6] Muhammad Tayyab Chaudhry, Teck Chaw Ling, Atif Manzoor, Syed Asad Hussain, and Jongwon Kim. 2015. Thermal-Aware Scheduling in Green Data Centers. ACM Computing Surveys 47, 3 (2015), 1-48. DOI: https://doi.org/10.1145/2678278

[7] Xibo Jin, Fa Zhang, Athanasios V. Vasilakos, and Zhiyong Liu. 2016. Green Data Centers: A Survey, Perspectives, and Future Directions. arXiv (August 2016). DOI: https://doi.org/ https://doi.org/10.48550/arXiv.1608.00687

[8] Greenpeace International Greenpeace International. 2018. How dirty is your data? (June 2018). Retrieved May 28, 2022 from https://www.greenpeace.org/international/publication/7196/how-dirty-is-your-data/

[9] Marta Chinnici, Alfonso Capozzoli, and Gianluca Serale. 2016. Measuring energy efficiency in data centers. Pervasive Computing (2016), 299-351. DOI: https://doi.org/10.1016/b978-0-12-803663-1.00010-3

[10] Alfonso Capozzoli, Marta Chinnici, Marco Perino, and Gianluca Serale. 2015. Review on Performance Metrics for Energy Efficiency in Data Center: The Role of Thermal Management. Energy Efficient Data Centers (2015), 135-151. DOI: https://doi.org/10.1007/978-3-319-15786-3_9

[11] Yeo, Sungkap & Hossain, Mohammad Mosaddek & Huang, Jen-Cheng & Lee, Hsien-Hsin. (2014). ATAC: Ambient Temperature-Aware Capping for Power Efficient Datacenters. Proceedings of the 5th ACM Symposium on Cloud Computing, SOCC 2014. 10.1145/2670979.2670996

[12] Lizhe Wang, Samee U. Khan, and Jai Dayal. 2011. Thermal aware workload placement with task-temperature profiles in a data center. The Journal of Supercomputing 61, 3 (2011), 780-803. DOI: https://doi.org/10.1007/s11227-011-0635-z

[13] SeyedMorteza Mirhoseini Nejad, Hosein Moazamigoodarzi, Ghada Badawy, and Douglas G. Down. 2020. Joint data center cooling and workload management: A thermal-aware approach. Future Generation Computer Systems 104, (2020), 174-186. DOI: https://doi.org/10.1016/j.future.2019.10.040

[14] Luca Parolini, Bruno Sinopoli, Bruce H. Krogh, and Zhikui Wang. 2012. A Cyber–Physical Systems Approach to Data Center Modeling and Control for Energy Efficiency. Proceedings of the IEEE 100, 1 (2012), 254-268. DOI: https://doi.org/10.1109/jproc.2011.2161244

[15] Equipment Thermal Guidelines for Data Processing Environments. Retrieved May 28, 2022 from https://www.ashrae.org/File%20Library/Technical%20Resources/Bookstore/datacom1_4th/ReferenceCard_7-25-16.pdf

[16] Ana Azevedo and M.F. Santos. KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW Ana. Retrieved May 28, 2022 from https://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf

[17] Zach (2021). When to Use Mean vs. Median (With Examples). [online] Statology. Available at: https://www.statology.org/when-to-use-mean-vs-median/.

[18] Patrick Schober, Christa Boer, and Lothar A. Schwarte. 2018. Correlation Coefficients. Anesthesia & Analgesia 126, 5 (2018), 1763-1768. DOI: https://doi.org/10.1213/ane.0000000000002864

[19] CRISP-DM - Data Science Process Alliance. Retrieved May 28, 2022 from https://www.datascience-pm.com/crisp-dm-2/

[20] Introduction to SEMMA - SAS Help Center. Retrieved May 28, 2022 from https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

[21] Anastasia Grishina, Marta Chinnici, Ah-Lian Kor, Davide De Chiara, Guido Guarnieri, Eric Rondeau, and Jean Philippe Georges. 2022. Thermal awareness to enhance data center energy efficiency. Cleaner Engineering and Technology 6, (2022), 100409. DOI: https://doi.org/10.1016/j.clet.2022.100409

[22] Davide De Chiara, Marta Chinnici, and Ah-Lian Kor. 2020. Data Mining for Big Dataset-Related Thermal Analysis of High-Performance Computing (HPC) Data Center. Lecture Notes in Computer Science (2020), 367-381. DOI: https://doi.org/10.1007/978-3-030-50436-6_27