Characterization and Forecasting the Workload of DRLAB Medical Database Server Based on the Shift-Wise and Machine Learning Approaches

Saleem Razzak Qureshi[*], Aftab Ahmed Chandio[†], Qamar-ul-Nisa Chandio[‡]

Abstract

The fundamental process to understand systems' workload is to analyze its impact and outlining workload characterization for better understanding, it enables systems' owners and policy makers to make decisions regarding policy management in order to improve system performance. Digital healthcare system is being modernized very fast at global level, especially medical laboratories require more advancements to match with standardization therefore efficient workload management of medical servers is needed to ensure reliable performance and scalability. This research focuses on workload characterization of medical database web servers, specifically within Diagnostic and Research Laboratory (DRLAB) at Liaquat University of Medical and Health Sciences (LUMHS), Jamshoro, Sindh, Pakistan. The laboratory runs completely on ICT-based infrastructure. It offers lots of end-users to connect with, like patients, lab technicians, doctors, administrative staff, IT persons, and office workers, which raises concerns about systems' scalability as end-users' requests vary throughout the day. To assess the load of end-users on servers' performance the approach is being used in this study is to analyze server access log data collected over seven-day period (4th to10th September 2020), comprising over 160,000 requests. We broke down the information into four six-hour shifts: midnight (12:01 AM to 6:00 AM), morning (6:01 AM to 12:00 PM), noon (12:01 PM to 6:00 PM), and evening (6:01 PM to 12:00 AM). In this way, the status of different time intervals in aspect to rush time may be observed. Furthermore, based of four observations, the machine learning Regression analysis techniques applied to comparatively analysis. Moreover, the results will be helpful to scheme up the database performance policy. The policy makers/stakeholders mitigate the issue by figuring out the analyzed data/statistics for the future planning.

***Keywords*:** Characterization, Database Workload, Machine Learning, Regression Analysis, Medical Lab, ICT, LUMHS, COVID.

Introduction

This particular research is carried out to characterize the workload of Medical Database Servers in order to evaluate system's performance. The workload characterization of a system is a basic process to describe

[*]Institute of Mathematics and Computer Science, University of Sindh Jamshoro, Jamshoro 76080, Pakistan, saleemrazzak786@hotmail.com

[†]Corresponding Author: Institute of Mathematics and Computer Science, University of Sindh Jamshoro, Jamshoro 76080, Pakistan, chandio.aftab@usindh.edu.pk

[‡] Government Degree Boys College Qasimabad, Education & Literacy Department, Government of Sindh, Hyderabad 71000, Pakistan, qamaraftabchandio@gmail.com

the systems' workload as the system's owner can take the decision for policy management to enhance the system performance.

This study targets to the medical investigation center called the Diagnostic and Research Laboratory (DRLAB) at Liaquat University of Medical and Health Sciences, Jamshoro (LUMHS). LUMHS is very old public sector Medical University established as a Medical School in 1981 and later on became a Medical University in the year 2001. Diagnostic and Research Laboratory started functioning at Civil Hospital Hyderabad Branch in 2002. Currently, this Laboratory has started its branches scattered across Sindh Province of Pakistan. The laboratory has been fully functioned through online based on Information Communication Technology (ICT) infrastructure. Since the laboratory system provides several services to a huge number of users (i.e., computer operators, patients, medical consultants, etc.), it is necessary to balance the workload and to optimize the system's performance (Chandio et al., 2014). The Laboratory is used to perform COVID-19 tests in Sindh Province in this regard providing online/live results access to stakeholders like health management departments of the province.

The purpose of this study is to focus on analyzing the impact due to increase in user(s) workloads, as that the policy makers/stakeholders mitigate the issue by figuring out the analyzed data/statistics for the future planning (Jhatial & Chandio, 2023). Medical Laboratory Database provides the data to the dedicated Portals for accessing COVID-19 Data, General Patients Reports, Investigation data for Research purpose.

This study uses a real-world workload of a medical database server consist of complex and large-scale architecture. The initial workload is collected from the DRLAB at LUMHS (https://drlab.lumhs.edu.pk/onlinereporting.php). The seven-day workload of a day (from 4th to10th September 2020) is distributed in four categories: (a) Mid-night (12:01 AM to 6:00 AM), (b) Morning (6:01 AM to 12:00 AM) (c) Noon (12:01 PM to 06:00 PM) (d) Evening (6:01 PM to 12:00 PM).

To maintain and tune-up the performance of any computerized system within complex and large-scale architectures, the workload characterization and analysis play a key role in the process (Calzarossa et al., 2016). It is impossible to analyze, classify, and summarize manually the data because of the incredible increase in data (Li & Beaubouef, 2010). It has been proofed in state-of-the-art that the best approach for analyzing workload is through statistical tools by providing the data with different time periods. The results will be helpful to scheme up the database performance policy.

This research is concerned about the scalability of database in case the number of users increase periodically and the workload impacts the database performance due to those users on different time periods. This study is to focus on analyzing the impact due to increase in users/user workloads so that the policy makers/stakeholders mitigate the issue by figuring out the analyzed data/statistics for the future planning.

In the organization of this paper, we describe related work in Section II, while Section III explains the methodology, system life-cycle architecture, dataset, tools, regression models, Section IV shows results, Section V describes the discussion.

Literature Review

In this section, we address the current state-of-the-art in the domain (Chandio et al., 2014; Jhatial & Chandio, 2023; Karmani et al., 2018; Khattak et al., 2025). Yu et al. (2017) studied Metric Importance Analysis for Big Data Workload Characterization, and mentioned that workload analytics is base of every business in this world of technology and workload runs on large-scale interconnected clusters can be understood by characterizing overall system performance (Yu et al., 2017). Impact on the databases due to increasing load of users has always been vital and day to day monitoring area (Raza et al., 2019). Saverimoutou et al. (2019) analyzed the Data of Access log (from April 2018 to April 2019), and used Statistical analysis techniques (Mean Distribution) in the Webview tool. Log file is belonged to Web server workload. Song and Mahanti (2019) assessed the Data of Access log (from 2012 to 2015), and used statistical analysis technique (Zapf model, Maximum likelihood, Weibull model, Lognormal model) in R-tool. The log file is belonged to Web server workload. Workload is differed from mobile and fixed device requests. Mobile requests had higher success rate and users request less web files, more images. Presented statistical models for mobile workload characteristics which can be used to improve existing traffic models. Xu et al. (2018) accessed Dataset of Access log (for one month), used Expectation-Maximization-based Modeling, the used Web server workload. Propose an efficient detection approach of web bot traffic to a large e-commerce marketplace. Samad et al. (2018) analyzed Data of Access log, and used LoadUIWeb , Webserver Stress tools. The dataset is about Web server workload. They found Web page attributes effect response times the most, and to speeding up web page response times are minimizing the number of embedded objects.

Jarkad and Bhonsle (2015) used the Accessed log of four days in January 2017, they used web usage mining techniques in the web Expert tool. While the log file is also from Web server workload to find user

behavioral pattern. Summers et al. (2016) accessed log file of 24 hours, in March of 2014, and used Prefetch Algorithm technique by simulation model. In this log file Video Service Workload is targeted. They analyzed the usage hard drives and system memory that holds perfected content.

Eldin et al. (2014) studied that Data log accessed from May,2008 to October, 2013 from web server workload, and the techniques for prediction such as Time-series analysis, ARIMA Model, polynomial splines are used in tool of R forecasting package. In the study of Jarkad and Bhonsle (2015), Dataset is a log file of web server workload and used web usage mining, backtracking algorithm, graph partitioning algorithm techniques. They Presented user navigation pattern prediction system which predict user future request in less time. To analyze users' impact while login to database and connection workload throughout the session is a key target of this research study (Korkmaz et al., 2018).

To better utilize the Physical Machines (PM) the workload prediction is very important, in case the monitoring and control of the workload on database systems is vital, the Prediction of workload is playing an important role in order to evaluate the performance and tune-up the database servers (Rossi et al., 2025).

Basically, in this study, a medical investigation center called the Diagnostic and Research Laboratory at LUMHS is targeted. The Diagnostic and Research Laboratory is started its functions firstly at Civil Hospital Hyderabad Branch in 2002. Now days, this Laboratory established its branches spread out across Sindh Province of Pakistan. The laboratory has been fully functioned through online based on ICT infrastructure. Since the laboratory system provides several services to a huge number of users (i.e., computer operators, patients, medical consultants, etc), it is necessary to balance the workload and to optimize the system's performance (Shishira et al., 2017). This research proposes a statistical and data mining technique to predict the future workloads in order to share with stockholders.

The quantitative performance evaluation is a major concern of computer science research and systems, especially for complex architectural software such as Database Management System (DBMS) (Zhang et al., 2017). Most modern systems generate abundant and diverse log data. With dwindling storage costs, there are fewer reasons to summarize or discard data. However, the lack of tools to efficiently store and cross-correlate heterogeneous datasets makes it tedious to mine the data for analytic insights (Bitincka et al., 2010). Performance evaluation of DBMS is a major issue since it is generally difficult to model experimental and performance analysis results (Raza et al., 2019; Zhang et al., 2017).

Impact on the databases due to increasing load of users has always been vital and day to day monitoring area (Raza et al., 2019). To analyze users impact while login to database and connection workload throughout the session is a key target of this research study (Korkmaz et al., 2018).

Methodology

In this section, we describe the life-cycle of the DRLAB system architecture, regression model, research tools, dataset, performance metrics.

Understating the Life Cycle of the DRLAB System

This research work is specifically carried out on the Diagnostic & Research Laboratory, LUMHS, Jamshoro/Hyderabad. Technically the database of this laboratory has the following nature of connections and its life-cycle is depicted in Figure 1. The life-cycle of the DRLAB can be described into three major phases: (a) Host machine interfacing phase; (b) Expert Laboratory information phase; (c) Online reporting phase.

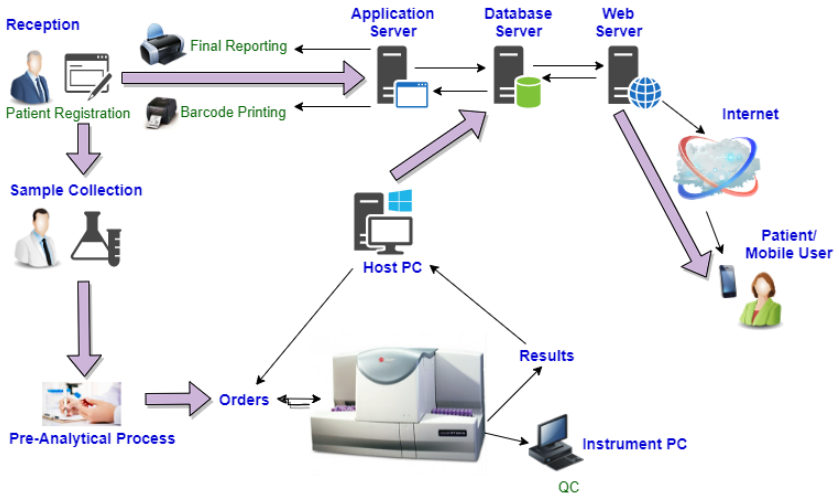

Figure 1: The life-cycle of DRLAB Medical Database Server System Architecture.

Machine Host Interfacing Phase

In the Laboratory, medical test instruments are connected with serial cables to Personal Computers (Clients), Host interfacing software are running on each client and after processing the test samples, instruments send the data to Host Interfacing Software. In order to realize the nature of automation system, every host interfacing software

establishes a connection to database to transmit the results for finalizing the reports for patients.

Expert Laboratory Information System (E-LIS) Phase for User Connections

An Expert-Laboratory Information System initiates database connection with login of reception users where patients first come to register their information by booking an order to perform required medical investigations. After this process the samples collected and forwarded to their corresponding investigation department to produce the results for reporting. On the other hand, those corresponding department's user are connected with database to processed the final results so that the patient can collect the report.

Online Reporting Phase for Patient's Login Sessions

Beside the above phases, an alternative way to get the investigation reports is also designed for online reporting web portal. By this portal the patient does not need to physical visit the nearest branch for collecting the reports, however, Patient has to login with Login ID and password mentioned on the Patient booking order slip and can see the results online on their internet devices.

In this way the overall Information and Automation System of the Diagnostic and Research Laboratory require connections to database and perform the operations for working cycle. So therefore, it is very important to analyze and monitor the complete system periodically and timely so that the system can be protected against failure due to lack of scalability.

Regression Model

Regression and statistical analysis approaches are used to analyze correlations between various performance measures and workload measures. The basic statistics methods for further analysis are also applied to establish trends, correlation, and predictive relationships to be used to predict system behavior. The results are used as the basis to evaluate system performance and to identify plans to enhance scalability and responsiveness.

Regression analysis is a standard statistical technique that helps determine the strength and direction of the relationship between two or more variables. Regression, specifically, is applied in analyzing the impact of the change in one or more independent variables (predictors) on a dependent variable (response). Apart from relationship analysis, regression is a convenient tool for forecasting and prediction from past trends.

For the purposes of this research, we utilized simple linear regression, which describes the relationship between one independent and a dependent variable in terms of a linear equation. The general mathematical formula of the linear regression model is presented as:

y=α+βxy = \alpha + \beta xy=α+βx

Where:

yyy = Dependent variable (response variable)

xxx = Independent variable (predictor variable)

α\alphaα = Intercept (the expected value of yyy when x=0x = 0x=0)

β\betaβ = Slope (rate of change in yyy for a unit change in xxx)

Research Tools

Statistical tools including R-tool, Stata, Microsoft Excel, Oracle Enterprise Manager, Trace Files, are considered to analyze the workloads. The studied workload is chosen according to network-intensive and data-intensive, which will be better choice to analyze the performance of the current system. Furthermore, we used the well-known data mining tool i.e., Weka for using data mining technique to predict the future workload (Garner, 1995).

Real-World Workload (Dataset) of Medical Database

This study uses a real-world workload of a medical database server consist of complex and large-scale architecture. The workloads are collected from the Diagnostic and Research Laboratory at LUMHS. The dataset is from seven-day period (4th to10th September 2020), comprising over 160,000 requests. The workload of a day is distributed in four categories: (a) Mid-night, (b) Morning (c) Noon (d) Evening.

Four time-based operational shifts are divided on bases of data, each lasting six hours. They are:

Night Shift (12:01 AM – 6:00 AM)

Morning Shift (6:01 AM – 12:00 PM)

Noon Shift (12:01 PM - 6:00 PM)

Evening Shift (6:01 PM – 12:00 AM)

The execution jobs in the workload are executed on the Physical machine(s) (PM), i.e., Intel Xeon E5-2650 v2. The PM is comprised of total 8 cores (16 threads processors) with 2.60 GHz speed and 80 GB memory.

Performance Evaluation Metrics

In this study, for statistical analysis, we considered the basic statistics functions including count, maximum, minimum functions. The total workload in terms of number of data size transferred calculated by

count function with the time intervals such as shift-wise six-hourly and hourly. The same method is also applied for calculating the workload in terms of hits or accessibly by the users with the time intervals such as shift-wise six-hourly and hourly.

As for as concerned the evaluation of the performance of regression analysis, we considered well-known performance metric is called *r-squired*. The *r-squired* (coefficient of determination) is a statistical measure in regression that indicates the percentage of variance in the dependent variable explained by independent variables. Ranging from 0 to 1 (0%–100%), it measures goodness-of-fit. How closely data fits the model. A higher indicates better fit

Results

In this section, we provide the results taken from our experimental work divided into two parts; (a) statistical analysis and (b) regression analysis.

Statistical Analysis

The analysis is focused on significant HTTP-based interactions captured in the dataset and helps to establish knowledge about system behavior under various circumstances. Specifically, the section examines the distribution and HTTP request method (i.e., GET, POST), HTTP response status code (i.e., 200 OK, 404 Not Found, 500 Internal Server Error), and HTTP protocol version (i.e., HTTP/1.1 or HTTP/2) used in client-server communication. Moreover, this provides an in-depth explanation of the data transfer statistics, for example, bytes received and sent per session. The statistics help determine the efficiency of the system and resource usage in processing user requests.

Firstly Figure 2 shows overall top ten clients, requested workload in aspect to number of hits. In this figure, x coordinate shows number of distinct clients, on the other side y coordinate shows total workload in aspect to hits. It can be clearly seen in the figure that client number 1 accessed the server more than 2000 times and rest of the clients between 1000 to 500 times. The client 1 may have very much frequently access to the database server due to its availability in centralized position in the branch circles.

In the chart, Figure 3 shows the cumulative number of clients versus the workload it produced, in terms of users. The number of clients is plotted on the x-axis and the total workload in terms of server hits on the y-axis. It can be observed from the graph that 148 clients accessed the server over 500 times, and on the other hand 250 clients (graph between 148 to 393 clients) accessed the server close to 100 times and the rest of

the clients made less than 50 hits. Furthermore, a graph of workload distribution by daily workload is depicted in Figure 4, plotting usage patterns over a number of days in terms of two important measures: data volume (kilobytes) and hits (number of hits). The x-axis of the graph is the number of days in the data set, and the y-axis is the corresponding workload in kilobytes.

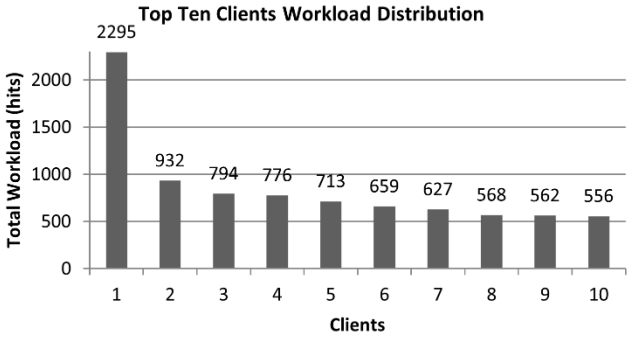

Figure 2: Top Ten Clients workload distribution.

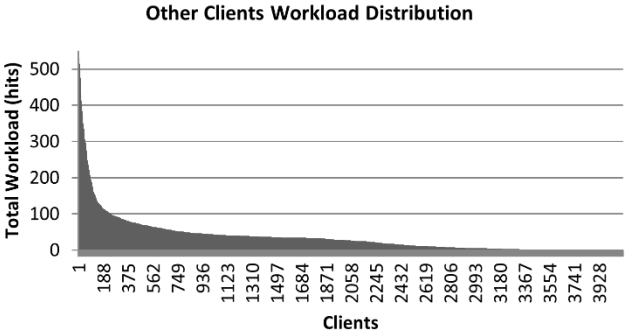

Figure 3: Other clients workload distribution.

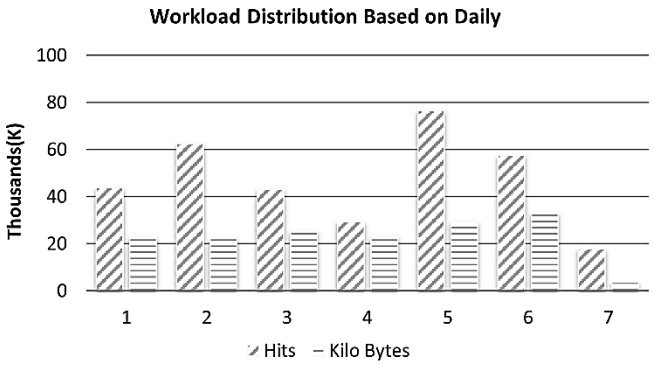

Figure 4: Workload distribution based on daily.

The chart employs two types of bar markers to identify between the readings: The series of straight-line bars show the volume of data (in kilobytes) consumed by the system on a day-to-day basis. The vertical-lines bars show the hits, i.e., the number of distinct client requests to the web/database server per day.

By visual observation of the chart, we can see that there is a sharp spike in activity within the system on the fifth day (8th September 2020) of the period of observation. Precisely, the consumption of data is more than 70 kilobytes, and the hits are over 30,000. The jump in an activity might mean we have hit a peak in how much the system is being used during user hit or that a lot more people are online at the same time. This analysis provides assumption of how busy the system gets each day to helps administrators and researchers to manage, detect anything unusual and figure out how much resources they need for better performance.

This sharp rise could indicate a peak usage incident or greater density of concurrent users, which might be an event of interest to analyze the cause behind, i.e., system update, reporting due date, or a public health-related peak in questionnaires. Overall, the figure provides valuable information on the trends of the system's daily workload, which would enable administrators and researchers to monitor intensity of use, detect anomalies, and assess resource needs better.

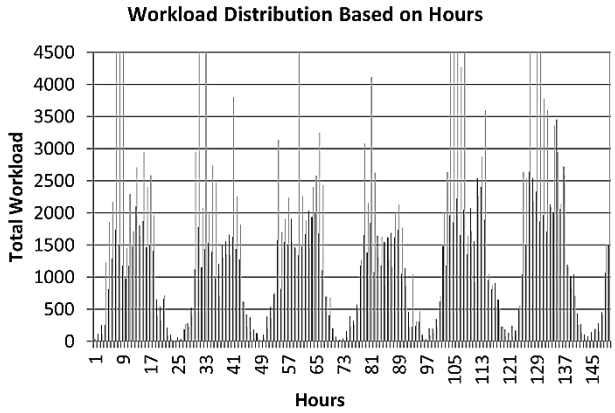

Figure 5: workload distribution based on hours (light color bar: total sum of data (KB), dark color bar: total no of hits in hour).

After analysis based on daily-wise, in this section a seven-day workload distribution of the medical database server based on hourly is described by example, as shown by Figure 5. The figure shows a graphical representation of the hourly segmented workload, where both the amount of data transferred and server hits are shown over a 24-hour period.

The x-axis of the chart is the hour of the day, from 00:00 (midnight) to 23:00 (11 PM). The Y-axis is the total workload, in thousands, to fit the enormous amount of data and requests. Two different measures are charted with color-coded bars: The light color bars indicate how much data is used, in kilobytes (KB), per hour. The dark color bars represent the total number of server hits, or user requests, which have been recorded in each corresponding hour. Both readings, from the graph, reflect intense usage during peak hours. Specifically, a number of hourly time slots reflect over 4500 KB usage of data, and the hits are over 2000 requests per hour.

Basically, these kinds of trends show that the system is being used a lot, and it is mandatory to find-out patterns of user behavior and sudden spikes of system. This visual presentation is useful in providing understanding of temporal usage trends to assist system administrators in recognizing particular event when the server is under maximum load. This process is important in workload distribution, resource management and optimization especially in data-intensive settings like medical laboratories.

This part shows a graphical view of medical database servers' distribution of daily tasks, as seen in Figure 6. The x-axis of the chart is graphed on the four shifts in sequence, while the y-axis represents the amount of hit requests on the servers, measured in thousands. The visual display clearly shows that the morning and noon shifts always receive constantly higher traffic, each exceeding 13,000 hits. It shows the fact that most database activity and user actions such as data input, report generation, and test result processing, for instance—are taking place within regular business hours. While on the other hand, the evening and night shifts show relatively reduces traffic volumes, indicating a lower system usage rate during these hours. This is due to reduction in personnel operations, a smaller number of test request inputs, and generally decreased user activity during late hours. These user patterns are essential to calculate maximum usage time and potential points of system resource overload. Understanding these patterns can facilitate resource optimization, balancing server loads, and maintenance task scheduling, thereby improving the overall throughput and reliability of the system.

Figure 7 shows the daily distribution workload of the medical database servers measured by system hits during four distinct periods of time. The data-set has been divided into four six-hour working shifts: Night (12:01 AM – 6:00 AM), Morning (6:01 AM – 12:00 PM), Noon (12:01 PM – 6:00 PM), and Evening (6:01 PM – 12:00 AM). In the graph x-axis demonstrates these time intervals, and on other side the y-axis shows the total hits the system gets measuring in thousands. This graph offers a clear comparative view of system activity throughout different

times of the day. Analyzing the data reveals that the morning shift carries the higher workload, as compared to rest of the shifts, indicating greater system usage and user activity during this period. The night, noon, and evening shifts, conversely, have relatively lighter workloads, each receiving fewer hits in the overall daily count. This outcome indicates that the system is experiencing peak operating workload during the morning shift, which can be caused by high user activity, data input, diagnostic tasks, or patient requests during normal working hours. These results are of prime significance to guide resource deployment, load balancing, and performance tuning efforts for optimal operation of the system during peak usage.

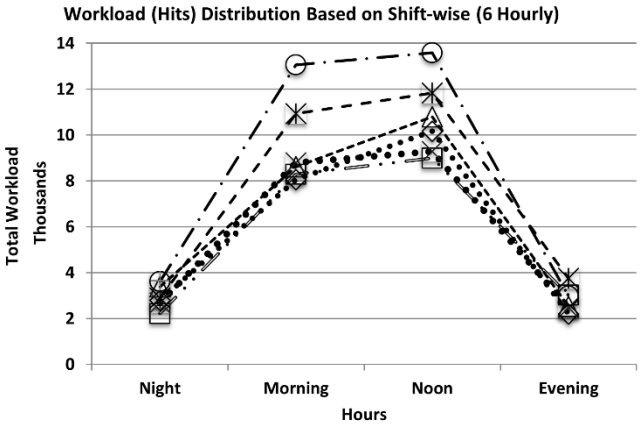

Figure 6: workload distribution-based hits shift wise.

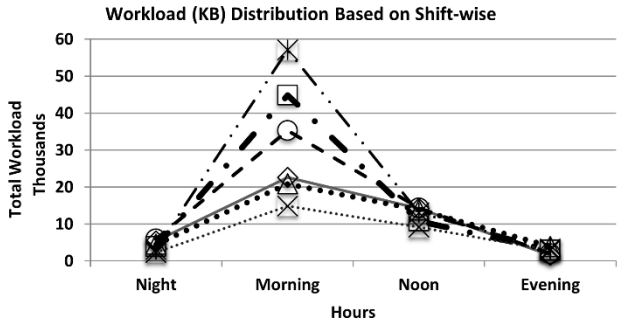

Figure 7: workload distribution-based kilobytes shift wise.

Here, Figure 8 depicts the distribution pattern of the workload throughout the day, in terms of the quantity of data transferred (in bytes). The X-axis of the graph is the server hits or client requests, and the Y-axis is the workload, measured in terms of the quantity of bytes processed. The

graph easily illustrates a pattern in the behavior of the system's data transfer. By far the overwhelming majority of requests are for data transfers less than 1,000 bytes, and this suggests that the majority of client-server traffic is lightweight transactions, perhaps in the retrieval of small records or meta-data. Few requests are crossing into the 12,000-byte range, and this suggests that heavy-duty data transfers are not standard and are probably tied to particular goals like report downloading, image retrievals, or complicated query results. This trend explains that when the system is handling a high volume of requests, the average request payload per request is very small, and this affects the network usage, server load, and response optimization. Identifying these trends is crucial for resource planning and performance tuning since it allows one to differentiate usual traffic from unusual load conditions.

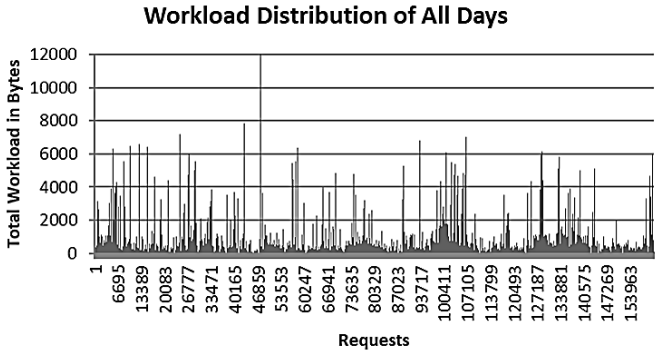

Figure 8: Workload distribution of all day.

Based on the analysis of the dataset, in Table 1, there were 160,634 client-server request communications. The requests were categorized based on the HTTP method, and it was observed that an overwhelming majority—150,353 requests—were of the GET type, which is used most prominently for retrieving resources from the server. The dataset also contained 10,218 POST requests, which are used most prominently for submission of data operations, i.e., form submissions or server updates. A further analysis of the dataset revealed that the requests were made by 4,120 unique client identifiers, which suggests a diverse set of users accessing the system within the observation time. The diversity suggests widespread use of the system and suggests needs for scalable server infrastructure to support diverse patterns of client activity.

Regression Analysis

In this section, we provide the results found after applying regression model. As we have distributed the dataset based on shift-wise

(i.e., described in dataset information section), the regression analysis is also applied on the four different observations. Their results are depicted in the Figures 9, 10, 11, and 12.

Table 1: Workload Detailed by HTTP Response Code.

Date	4-Sep-2020 to 10-Sep-2020
Total Requests	160634
GET	150353
HEAD	63
POST	10218
HTTP Response Status Codes	
Code	Hits
200	150321
206	1476
301	652
302	7091
304	905
400	2
404	5
500	182
Total Clients	4120

Observation 1

In this study, the observation 1, illustrates in Figure 9, deals with specific time period in the following way: During the Noon shift, the amount of data transferred (in kilobytes) increases with the ratio of the number of hits to the website. To validate this observation, linear regression analysis was performed. In this, independent variable is the number of kilobytes transferred, and the dependent variable is the number of hits on the site.

Observation 2

In the second observation, volume of Data against Website Hits during the Morning Shift is observed in the Figure 10. During the morning shift, the trend was seen where the volume of data transferred, in kilobytes (KB), rose proportionally with the number of hits on the site. The independent variable (predictor) used in Figure 10 is the volume of data received in kilobytes, while the dependent variable (response) is the volume of hit received. The aim of regression analysis is to determine whether fluctuations in data volume can be used to predict a change in the volume of hits received with precision.

Observation 3

In the third observation, we considered number of Hits against Volume of Data During Evening Shift in Figure 11. Through examination of system performance over different time spans, Observation #3 reveals a very strong pattern developing in the evening shift. There is a clear

increase in the volume of data transferred (in kilobytes) that is accompanied by an increase in the number of hits. In order to explore this relationship more specifically, a linear regression analysis was performed using the evening-specific dataset. Two variables were selected in this analysis: The independent variable (predictor) is the number of kilobytes transferred. The dependent variable (response) is the hits recorded in the evening shift.

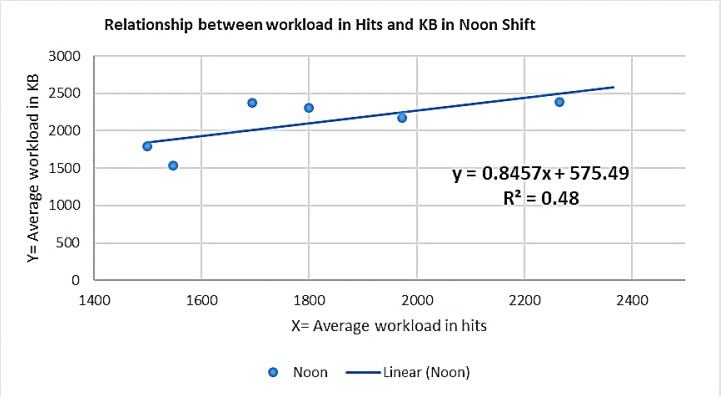

Figure 9: Relationship between kilo bytes and Requests (Hits) in Noon Shift.

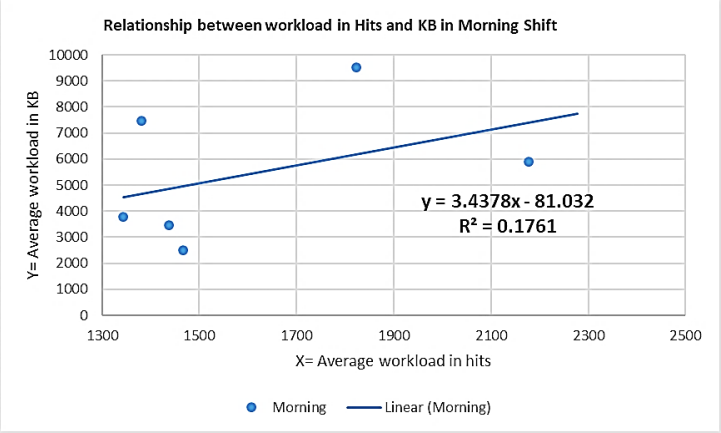

Figure 10: Relationship between Kilobytes and Requests (Hits) in Morning Shift.

Observation 4

In the last observation, the greater Volume of Data (Kilobytes) can be correlated with greater Hit frequency during Night Shift is observed in Figure 12. In the comparison of system activity over various time periods, there was one trend that was strong in the night shift data set, in which the

level of kilobytes transferred was correlated with an increase in the level of hits on the server. To further examine this correlation, a linear regression analysis was run based on the night shift data since there were two very distinct variables: the independent (predictor) variable, the level of the data in kilobytes, and the dependent (response) variable, the level of hits on the server. The reason for using linear regression here was to determine the nature and scale of the relationship between the volume of data transmitted and the quantity of user hits on the server during this specific time frame.

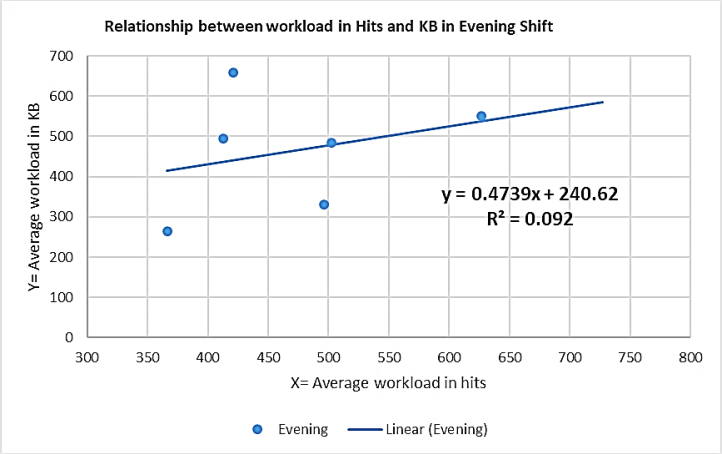

Figure 11: Relationship between Kilobytes and Requests (Hits) in Evening Shift.

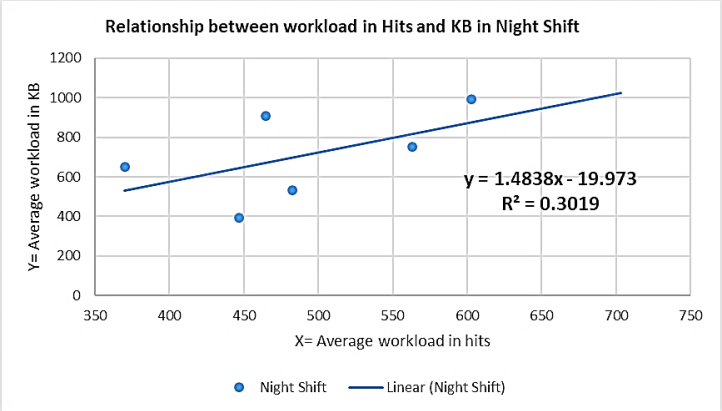

Figure 12: Relationship between Kilobytes and Requests (Hits) in Night Shift.

Discussion

In this section we provide the discussion on the results taken during statistical and regression analysis. The results of Figure 9, enables us to see if there exists a statistically significant relationship between the amount of data sent and the amount of user interaction. The results indicate a high positive correlation between the two variables. The test of regression confirms that as the Figure 9 for kilobytes increases, the figure for hits also increases proportionately. The positive linear correlation confirms the hypothesis that higher levels of data transference in the Noon shift are associated with greater web server utilization, reflecting an increase in user activity within this time frame. This is crucial to system administrators because it means that there must be resource optimization and load balancing during peak demand times in an effort to maintain the system.

The results of Figure 10 show a potential correlation between the rate of access and utilization of data within the specified time frame. The regression line, graphed in Figure 10, indicates that there is a weak positive relationship between the two variables. That is, as hits rise, so does the volume of data transfer, but slightly. The relationship is weak, though, which indicates that while there is some relationship, there are probably other variables at play that influence the volume of data as well and hits alone do not have a strong influence on the variation in kilobytes transferred. This result suggests a moderate but genuine dependence of user activity on data load during the morning shift, one that deserves further investigation of other variables—such as the nature of the content being queried or the nature of the user sessions—where there might be a more robust explanation of empirical data patterns.

The result of the regression of observation 3 is presented in Figure 11. That is, as the number of user requests (hits) increases, then there is a corresponding increase in the volume of data transferred by the server. The figure depicts a weak positive correlation, and this implies that as the amount of data increases, the server hits also rise slightly. The correlation is weak, and this indicates that although there is a positive relationship between the two variables, the relationship is not strong enough to establish a clear cause and effect. This weak positive trend may be explained through the characteristics of the content requested during the evening shift, which could entail a combination of regular and heavy data requests. It raises the demand for more in-depth study regarding the content types requested and the user behavior during this time frame, along with the server's response mechanisms under different intensities of workloads.

As Figure 12 illustrates, the result was that there was a weak but positive relationship between the two variables, which shows that the quantity of hits increases as the volume of data (in kilobytes) increases. But it should be pointed out that the correlation seen is very weak, and it is only modest linear association. This would imply that while a tendency exists for data volume to increase with user activity, other factors could be operating here as well, and the relationship is not strong enough to draw conclusions without further investigation.

From the literature review, we found in the state-of-the-art, that authors addressed the similar problem of this study and applied regression model for analysis of web and database workload via *r-squired* performance evaluation metric (Eldin et al., 2014; Jarkad & Bhonsle, 2015; Jhatial & Chandio, 2023; Saverimoutou et al., 2019). The shift-wise six-hourly and hourly distribution of the workload and its observations are only considered in this study. Additionally, though there are various built-in and third-party tools available in the market and were considered in the above studies but do not support customization and are very slow and that is considered in this study. The purpose of this study is to focus on analyzing the impact due to increase in user(s) workloads, as that the policy makers/stakeholders mitigate the issue by figuring out the analyzed data/statistics for the future planning (Jhatial & Chandio, 2023).

Conclusion

Basically, in this study, a medical investigation center called the Diagnostic and Research Laboratory at LUMHS is targeted. The laboratory has been fully functioned through online based on ICT infrastructure. Since the laboratory system provides several services to a huge number of users (i.e., computer operators, patients, medical consultants, etc.), it is necessary to balance the workload and to optimize the system's performance. It has been observed that the morning and noon shifts were the busiest one, which really affected how fast the servers responded and how much they were utilized. This research proposes a statistical and machine learning technique to predict the future workloads in order to share with stockholders.

Acknowledgment

Saleem Razzak's work was supported for his MPhil studies University of Sindh, Jamshoro, Pakistan. Web Log Dataset received from the LUMHS Jamshoro granted NOC approval to carry out the study (Ref: Letter No. LUMHS/D&R-LAB/04028/25).

References

Bitincka, L., Ganapathi, A., Sorkin, S., & Zhang, S. (2010). Optimizing data analysis with a semi-structured time series database. Workshop on Managing Systems via Log Analysis and Machine Learning Techniques (SLAML 10),

Calzarossa, M. C., Massari, L., & Tessera, D. (2016). Workload characterization: A survey revisited. *ACM Computing Surveys (CSUR)*, *48*(3), 1-43.

Chandio, A. A., Zhang, F., & Memon, T. D. (2014). Study on LBS for characterization and analysis of big data benchmarks. *Mehran University Research Journal of Engineering and Technology*, *33*(4), 432-440.

Eldin, A. A., Rezaie, A., Mehta, A., Razroev, S., de Sjöstedt-de Luna, S. S., Seleznjev, O.,…Elmroth, E. (2014). How will your workload look like in 6 years? analyzing wikimedia's workload. 2014 IEEE international conference on cloud engineering,

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. Proceedings of the New Zealand computer science research students conference,

Jarkad, P. M. P., & Bhonsle, M. (2015). Improved Web Prediction Algorithm Using Web Log Data. *International Journal of Innovative Research in Computer and Communication Engineering*, *3*(5).

Jhatial, S., & Chandio, A. (2023). Analysis of Academic Web Server Traffic and Workload Characterization for Performance Evaluation. *Sindh University Research Journal - SURJ (Science Series)*, *55*(02), 45-55.

Karmani, P., Chandio, A. A., Korejo, I. A., & Chandio, M. S. (2018). A review of machine learning for healthcare informatics specifically tuberculosis disease diagnostics. International conference on intelligent technologies and applications,

Khattak, E., Ullah, H., Khan, I., Ali, M. T., & Jan, L. (2025). Analyzing Twitter Data for Depression Signs Based on Machine Learning Techniques. *The Sciencetech*, *6*(4), 166-186.

Korkmaz, M., Karsten, M., Salem, K., & Salihoglu, S. (2018). Workload-aware CPU performance scaling for transactional database systems. Proceedings of the 2018 International Conference on Management of Data,

Li, Y., & Beaubouef, T. (2010). Data mining: concepts, background and methods of integrating uncertainty in data mining. *CCSC: SC Student E-Journal*, *3*, 2-7.

Raza, B., Sher, A., Afzal, S., Malik, A. K., Anjum, A., Kumar, Y. J., & Faheem, M. (2019). Autonomic workload performance tuning in large-scale data repositories. *Knowledge and Information Systems*, *61*(1), 27-63.

Rossi, A., Visentin, A., Carraro, D., Prestwich, S., & Brown, K. N. (2025). Forecasting workload in cloud computing: towards uncertainty-aware predictions and transfer learning. *Cluster computing*, *28*(4), 258.

Samad, H., Hanizan, S., Din, R., Murad, R., & Tahir, A. (2018). Performance evaluation of web application server based on request bit per second and transfer rate parameters. Journal of Physics: Conference Series,

Saverimoutou, A., Mathieu, B., & Vaton, S. (2019). Influence of internet protocols and CDN on web browsing. 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS),

Shishira, S., Kandasamy, A., & Chandrasekaran, K. (2017). Workload characterization: Survey of current approaches and research challenges. Proceedings of the 7th international conference on computer and communication technology,

Song, Y.-D., & Mahanti, A. (2019). Comparison of mobile and fixed device workloads in an academic web server. 2019 IEEE International Symposium on Measurements & Networking (M&N),

Summers, J., Brecht, T., Eager, D., & Gutarin, A. (2016). Characterizing the workload of a Netflix streaming video server. 2016 IEEE International Symposium on Workload Characterization (IISWC),

Xu, H., Li, Z., Chu, C., Chen, Y., Yang, Y., Lu, H.,…Stavrou, A. (2018). Detecting and characterizing web bot traffic in a large e-commerce marketplace. European Symposium on Research in Computer Security,

Yu, Z., Xiong, W., Eeckhout, L., Bei, Z., Mendelson, A., & Xu, C. (2017). Mia: Metric importance analysis for big data workload characterization. *IEEE Transactions on Parallel and Distributed Systems*, *29*(6), 1371-1384.

Zhang, M., Martin, P., Powley, W., & Chen, J. (2017). Workload management in database management systems: A taxonomy. *IEEE transactions on knowledge and data engineering*, *30*(7), 1386-1402.