A MACHINE LEARNING APPROACH TO MAINFRAME ANALYSIS

1P.Priyanka, 2Deivanai Kathiresan
1,2School of Computing Science and Engineering, Vellore Institute of Technology Chennai Campus, Chennai, India. 
1priyanka.p2015@vit.ac.in, 2deivanai.kathiresan@vit.ac.in

Abstract: Mainframe System processing includes a ‘Batch Cycle’ that approximately spans 8 pm to 8 am ,every week, from Monday night to Saturday morning. The core part of the cycle completes around 2 am, with key client deliverables associated with the end times of certain jobs, tracked by Service Delivery. There are single and multi-client batch streams, a QA stream which includes all clients, and about 2,00,000 batch jobs per day that execute. Despite a sophisticated job scheduling software, and automated system workload management, operator intervention is required, or believed to be required, to reprioritize when and what jobs get available system resources. Our work is to characterize, analyse and visualize the reasons for a manual change in the schedule. The work requires extensive data pre-processing and building machine learning models for the causal relationship between various system variables and the time of manual changes.

Keywords: Mainframe; Models; Commands.

1. INTRODUCTION

A centralized server is the thing that organizations use to have their business databases, exchange servers, and applications that require a more noteworthy level of security and accessibility than is usually found on litter scale machines. Centralized computer dependably Contain around seventy percent of corporate information from operations (bookkeeping, finance, charging, etc.) Often the "database server" in web-empowered database applications

Incorporated PCs will be PCs used chiefly by immense relationship for essential applications, conventionally mass data planning. Present day centralized server PCs have capacities less characterized by their single assignment computational speed (tumbles or clock rate) as by their repetitive interior designing and coming about high dependability and security, broad info yield offices, strict in reverse similarity for more seasoned programming, and high use rates to bolster enormous throughput. These machines regularly keep running for a considerable length of time without intrusion, with repairs and even programming and equipment updates occurring amid ordinary operation. For instance, ENIAC stayed in consistent operation from 1947 to 1955. All the more of late, there are a few IBM centralized server establishments that have conveyed over 10 years of ceaseless business benefit starting 2007, with redesigns not intruding on administration. Centralized servers are characterized by high accessibility, one of the primary purposes behind their life span, as they are utilized as a part of uses where downtime would be expensive or disastrous. The term Reliability, Availability and Serviceability (RAS) is a characterizing normal for centralized computer PCs.

This kind of approach requires extensive data preprocessing and building machine learning models for the causal relationship between various system variables and the time of manual changes.

2. PROBLEM STATEMENT

To Characterize the state of the system using both business exceptions and system workload artefacts to determine if there are patterns in operator’s response, capturing this processing knowledge and if the type of manual intervention is (1) predictable and (2) Can be automated. Develop a real time decision support application based on learning how the system state changes and is related to. So that reliance on operator
experience to meet business goals can be reduced while continuing to maximize the use of available resources.

3. PROPOSED WORK
Firstly, this phase consists of extracting the relevant data from the text files and storing them into database on which various analysis would be performed. Generalizing CPU health data for the whole data by linearly extending the data based on fixing the intervals. Also by standardizing the entries into one format and store them as tables in a database. From those files we track the JES commands given by the operator which can be found from the references. Finally in this phase tables are created based on these commands and the relevant information is stored. The database contains all the definitions which have made for planning objects. It likewise holds insights of employment and occupation stream execution and in addition the data as the client Id who has made a protest and with that it indicates when the last question was modified. Upon this we build a machine learning model on top of this so that we can segment the jobs based on the priority and plan it accordingly which has to be executed.

4. METHOD OF APPROACH
4.1 MAINFRAME APPROACH
Centralized server preparing incorporates a 'group cycle' that roughly traverses 8 PM to 8 AM, every week from Monday night to Saturday morning. The center part of the cycle finishes around 2 AM with key customer deliverables connected with the final days of specific employments, followed by administration conveyance. There are single and multi-customer clump streams, a QA stream which incorporates all customers and around 2,00,000 bunch occupations for every day that executes. In addition to the day time business transaction there are also client and securities information vendor files receive as input into the streams. There is a relative job priority classification scheme. There is a job scheduling application to manage the submission of jobs, based on time and other job or file input delivery dependencies.

4.2 JOB SCHEDULING
During the batch cycle, the mainframe system run at or near 100% of capacity. Despite a sophisticated job scheduling software and automated system workload management, Operation intervention is required or believed to reprioritize when and what jobs get available system resources to ensure tracked deliverables and business expectations for the night are met. Operator experience with workloads and nightly business expectation variables results in changes to how the schedule executes. These changes are captured by manual operator interventions. Ideally the job schedule should execute on a daily basis without operator intervention, leaving the maximization of the system resources to the system workload manager, with all deliverables being met according to their relative priority for the business as defined in the job schedule. If that equilibrium condition is not being met, the signature of their departure from equilibrium are certain operator commands captured in the system log and certain types of messages in the scheduler log, signifying a state change.

4.3 TIVOLI WORKLOAD SCHEDULER (OPCA)
TWS1 is a fully automated batch job scheduling system that improves job throughput and greatly reduces operations. TWS helps you arrange and sort out each period of cluster employment execution. Amid the handling day TWS generation control programs deal with the creation control programs deal with the creation environment and computerize most administrator exercises it readying your employments for execution, resolves interdependencies and dispatches and tracks every occupation. Since your occupations start when their conditions are satisfied, idle time is minimized and throughput enhances essentially. Employments never come up short on arrangement and if a vocation falls flat, TWS handler the recuperation procedure with almost no administrator intercession.

4.4 JOB ENTRY SUBSYSTEM2
MVS (or Z/OC which is the working framework for IBM centralized servers) utilizes a vocation passage subsystem (JES) to get job1 into the working framework, plan occupations for preparing by mvs and control work yield processing. JES2 is plunged from HASP (Houston programmed spooling need) which is
characterized as a PC program that gives supplementary employment administration capacities, for example, Scheduling, control of employment stream and spooling. JES2 is an utilitarian expansion of the HASP program that gets occupations into the framework and process all yield information created by the occupation. JES2 is the part of mvs that gives the essential capacity to land positions into and yield out of the ms framework. It is intended to give effective spooling, planning and administration offices for the mvs working framework.

MCP commands are for the scheduler (TWS/OPCA) and JES commands are for OS (IBM Z/OS). MCP commands can get the jobs into the queue including changing its priority and service class. You can even remove the job from the queue using MCP. But one initiators pick up a jobs from the queue using MCP, if the initiator pick up a job then MCP commands cannot reach them only JES commands can reach.

4.5 ALGORITHM
• DECISION TREE

Decision tree can be developed generally quick contrasted with different techniques for characterization explanations can be built from tree that can be utilized to get to databases effectively. Refer Fig1. Below which shows the Decision tree classifiers acquire comparative or better precision when contrasted and other grouping strategies.

Figure 1: Decision Tree

Various information mining methods have as of now been done on instructive information mining to enhance the execution of understudies like Regression, Genetic calculation, Bays order, k-implies grouping, relate rules, expectation and so on. Information mining methods can be utilized as a part of instructive field to improve our comprehension of learning procedure to concentrate on recognizing, extricating and assessing factors identified with the learning procedure of understudies. Grouping is a standout amongst the most as often as possible. The C4.5, ID3, CART decision tree are connected on the information of understudies to foresee their execution.

• CART ALGORITHM

CART is defined as Classification And Regression Trees. The order tree development via CART depends on paired part of the traits. CART additionally in light of Hunt’s calculation and can be actualized serially. Gini list is utilized as part measure as a part of selecting the part quality. CART is unique in relation to other Hunt's based calculation since it is additionally use for relapse examination with the assistance of the relapse trees. The relapse investigation highlight is utilized as a part of estimating a needy variable given an arrangement of indicator factors over a given timeframe. CART S bolsters constant and ostensible property information and have normal speed of handling.

• K-NEAREST NEIGHBOR ALGORITHM

Assume that a question is inspected with an arrangement of various characteristics, yet the gathering to which the protest has a place is obscure. Expecting its gathering can be resolved from its qualities; diverse calculations can be utilized to mechanize the grouping procedure. A nearest neighbor classifier is a framework for describing segments in perspective of the course of action of the segments in the arrangement set that are most similar to the experiment. Below Fig 2. which
shows this kind of technique we can get the closest nearing neighbours.

Figure 2: K-Nearest Neighbour

4.6 NAIVE BAYES
If the inputs are independent, we will be using Naive Bayes technique to solve the problem. Given a game plan of things, each of which has a place with a known class, and each of which has a known vector of components, our indicate is build up a lead which will allow us to dole out future articles to a class, given only the vectors of elements portraying the future things. Issues of this kind, called issues of directed request, are ubiquitous, and various techniques for building such standards have been delivered. One crucial one is the guileless Bayes procedure—in like manner called nitwit’s Bayes which depicts the Fig 3. Below called, essential Bayes, and self-rule Bayes. This method is basic for a couple reasons. It is definitely not hard to assemble, not requiring any convoluted iterative parameter estimation arranges. This infers it may be expeditiously associated with gigantic data sets. It is definitely not hard to interpret, so customers clumsy in classifier development can grasp why it is making the portrayal it makes.

Figure 3: Naive Bayes

4.7 SUPPORT VECTOR MACHINES (SVM)
In support vector machine the data in plotted in the n-dimensional space. Fig 4. which shows After plotting the data in the n-dimensional space the data is splitted separating the different classes which involves method of supervised learning classification in the n-dimensional space. Based on that we will be drawing a line which is called as hyperplane since we are drawing the line in the n-dimensional space. After drawing the hyperplane we will seeing the classes which is having the highest margin. The classes which are best suited that is nearer to the hyperplane that is having more distance between nearest data point and the hyperplane.

Figure 4: Support Vector Machines

- THE APRIORI ALGORITHM
In apriori algorithm, if we consider n itemset then n set of rules are generated. Based on that among the n rules we need to find the rule which is having more support and confidence. For this we will using a best algorithm called Apriori algorithm which is refered to the Fig 5. Below. Firstly we need to generate frequent itemsets which is having more support and we need to change the rules with having more confidence based on the splitting of items.
5. LITERATURE SURVEY

Machine Learning gives capacity for projects to learn without being expressly customized for a specific dataset. Edmondson's insight is that ML is part of a software engineering thread known as model-driven engineering. ML introduces a new category of model-building activities that can transform the software development life cycle. ML is coming to a mainframe near you, but it may be cloaked in predictive analytics. Last year Zementis, whose products leverage the Predictive Model Markup Language (PMML), announced availability for z/OS. Zementis models can be used to embed predictive models in z/OS CICS or Web Sphere settings. The models are “write-once,” meaning they can be deployed to z/OS SPSS, R, Python, or SAS. In a post on IBM Developer Works, Ravi Kumar outlines how z/OS users can now enable ML on OLTP applications, such as by embedding predictive models in DB2. One technique embeds the z/OS SPSS Scoring Adapter for DB2. Another approach combines a PMML model with business rules to make real-time decisions in DB2 or use Zementis-generated PMML to inject in-app scoring for CICS or Java apps. The IBM DB2 Analytics Accelerator for z/OS (IDAA) supports several major predictive analytics algorithms: K-Means, Naive Bayes, Decision Tree, Regression Tree, and Two-step.

5. APPROACH FOR MODEL BUILDING

First Part: Data Extraction and Standardisation

This phase consists of extracting relevant data from the log files and storing them into the database on which various analysis would be performed. We generalise the data for the whole day by linearly extending the data in 15 min intervals. We track the entries in the log files which are of type MCP (Modify Current Plan) and JES2 (Job Entry Subsystem) commands. We standardise the entries into one format and store them as tables in a database. From the extracted data we are only interested in the following:

1. Date and time of occurrence
2. Application Number
3. Jobname
4. Jobnumber
5. Command specific details

Second Part: Machine Learning Model

The objective of this phase is to relate each of the 8 logical partitions taken at 15minutes time interval across the six system properties to the time the changes have taken place in the logical partitions. To analyse this various machine learning classification algorithms can be used such as logistic regression, decision tree, support vector machine etc.

6. ANALYSIS

The Fig6. Shows below the data is skewed and the line is in the increasing when compared to residual and leverage.
Figure 6: Logistic Regression

The Fig 7. shows the data is classified on the basis of decision tree for building a model.

Figure 7: Decision Tree

Figure 8: Residual Analysis

The Fig 8. which is residual analysis graph which is used for comparing how the data is skewed and how it is performing over the data.

7. CONCLUSION

By identifying the high priority jobs which are having higher wait time, making them to allocate first so that the higher priority jobs will gets executed first. Therefore we can characterize analyze and visualize the reasons for a manual change in the schedule.

REFERENCES


[13] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Elsevier,1992


