Rotterdam school of management

Time is Money: An Analysis of the Drivers of Cloud Performance When Hosting Credit Models

Author:
Benjamin F. Aston[a]

[a]Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands

Supervisor:
Jason Roos[b]

[b]Department of Marketing, Rotterdam School of Management, The Netherlands

February 27, 2025

Abstract

This article employs the use of causal learning models to investigate the underlying drivers of cloud performance. The main problem aimed to be solved is that many firms who use cloud computing resources are unaware of what drives the cloud performance and usage. After scouring current academic and managerial literature, a gap was identified whereby performance testing has never previously been done using causal learning techniques. This paper wishes to run a randomised experiment, assigning requests to random configurations, aiming to estimate the CATE through causal learning models. The research was undertaken in collaboration with a large multinational bank.

Keywords

Cloud Performance, Causal Inference, Bayesian Neural Networks, Causal Forest, Optimization.

Contents

1 Introduction

Time. Money. Both of these are scarce resources and often trade off against one another. In many contexts, entities attempt to maximise efficiency, often measured by these metrics [Qureshi et al., 2020]. Over the past three decades, we have seen the rise in cloud computing and its 'pay as you go' service [Li and Kumar, 2022], trading off performance and therefore time for a cost. The degree to which these concepts trade off is up to the entity in charge, but all aim for the highest marginal gain in utility. Cloud computing hosts an array of products, ranging from streaming videos to complex risk models. The question thus stands, where do we draw the line between optimising performance and cost? During this paper, an attempt at answering this question will be made, focusing on the case study of a large multinational bank, hereafter called, 'The Bank'.

The goal is to find the optimal balance of cost and time while remaining fast while benefiting from reduced costs. The Bank currently has no insight into the causal drivers of performance and tends to set up their system configurations based purely on rules of thumb with no empirical backing. To solve this problem, causal learning models will be employed so as to algorithmically uncover the causal drivers of performance, giving insight into future cloud requirements. The benefits are twofold. From one side, this will allow managers to throttle their usage up or down, while, on the other hand, academically, demonstrate an innovative application of recently developed methods to performance testing.

2 Problem statement and research question

Motivating the problem further, 'The Bank', much like many other cloud customers, tend to lack the understanding of the drivers of cloud performance when setting up their cloud configurations [Makhlouf, 2019]. This leads to volatile performance and high costs. The problem statement can therefore be stated as the following. "Cloud computing is costly and used by a plethora of customers, of which usage is often driven by unknown underlying configurable factors. These factors contribute to higher costs and poor performance." Building on this, the research question is therefore, "What are the causal drivers of cloud performance and what are the relationships between each of these drivers when influencing execution time?"

This paper adds value by uncovering the underlying drivers of cloud performance, a field few firms have tackled. Secondly, the methodology enacted during this project shall attempt to be a guide on how best to approach detailed performance testing from a causal learning perspective which has never been done before.

3 Research approach and experimental design

The experimental design has several key aspects, starting with the variables. The independent variables considered in this experiment include CPU Cores (amount of CPU cores in whole numbers), System Memory (amount of system memory, categorised into three buckets), MAS Workers (amount of MAS Workers in whole numbers) and

Database Memory (Amount of database memory, categorised into three buckets). The dependent variable is Execution time. Based on the research question, it is believed that the independent variables drive changes in the dependent variable. Furthermore, the independent variables are thought to interact with one another in influencing the execution time (dependent variable) [Wang et al., 2023]. This implies that the contribution of a one unit change in a independent variable will have a non linear influence on execution time assuming another independent variable has changed in value too. These points lead to the hypotheses.

Firstly, on the whole, this paper hypothesizes that the execution time is determined through a causal effect of both direct and indirect effects, attributed to the number of 'MAS Workers', 'CPU Cores', 'System memory' and 'Database memory'. The purpose is to estimate the conditional average treatment effect in the change in execution time which is believed to be none zero such that the configuration change influences execution time. This is represented by the null and alternative hypotheses shown in formulas 1 and 2.

$$H_0 : E[\tau(X)] = 0 \tag{1}$$

$$H_A : E[\tau(X)] \neq 0 \tag{2}$$

where:

- $E[\tau(X)]$ denotes the expected (mean) Conditional Average Treatment Effect (CATE) across all configurations where the configuration represents a row vector of all independent variables.

Next, it is hypothesized that each independent variable will interact with one another, such that each configuration (combination of independent variables) will have heterogeneous CATE estimates. This is represented by the null and alternative hypotheses in formulas 3 and 4.

$$H_0 : \tau(X_1) = \tau(X_2) = \cdots = \tau(X_n) \tag{3}$$

$$H_A : \exists i, j \quad \text{s.t.} \quad \tau(X_i) \neq \tau(X_j) \tag{4}$$

where:

- $\exists i, j$ represent two of the total configurations of "X"
- $\tau(X_1), \tau(X_2), \ldots, \tau(X_n)$ represent the CATE estimates for different cloud configurations denoted by "X".

Based on the aforementioned hypotheses, we can represent them with the following Directed acyclic graph (DAG).

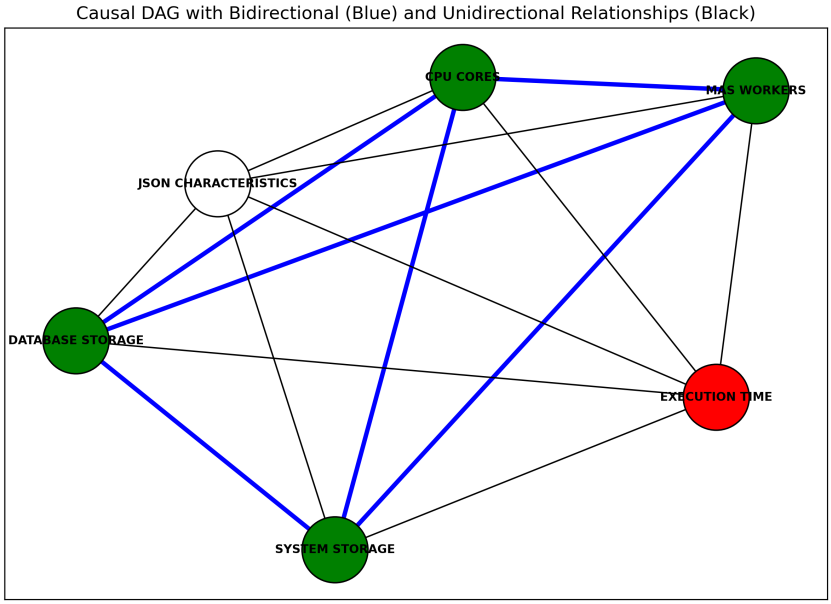

Figure 1: Causal DAG for Execution Time

Formally, these relationships are represented with equations five and six. In total, we expect 16 different beta estimates.

$$\text{Time} = Ck \tag{5}$$

where:

$$\begin{aligned} k = {} & \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \\ & + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{14} X_1 X_4 \\ & + \beta_{23} X_2 X_3 + \beta_{24} X_2 X_4 + \beta_{34} X_3 X_4 \\ & + \beta_{123} X_1 X_2 X_3 + \beta_{124} X_1 X_2 X_4 \\ & + \beta_{134} X_1 X_3 X_4 + \beta_{234} X_2 X_3 X_4 \\ & + \beta_{1234} X_1 X_2 X_3 X_4 \end{aligned} \tag{6}$$

where C is a scaling constant such that the relationship between K and Y scales linearly depending on the values of each independent variable that is used in the estimate of their beta coefficients.

where:

$$\begin{aligned} X_1 &= \text{CPU CORES} \\ X_2 &= \text{SYSTEM MEMORY} \\ X_3 &= \text{MAS WORKERS} \\ X_4 &= \text{DATABASE MEMORY} \end{aligned}$$

Time is therefore a function of all independent variables and their interactions for the given assigned system configuration. It must be stressed that these relationships are based purely on managerial and academic research. That said, this paper will use Order Space sampling to uncover the true edges of the DAG in an algorithmic fashion as shown by Ellis and Wong [Ellis and Wong, 2008]. The variables will remain the same,

but the best fitting DAG will be selected using Bayesian scoring methods during the analysis phase, redefining the DAG.

The experiment will be carried out through the means of a two stage, four arm trial whereby we have a cloud request that we then randomly assign to 25% of the cloud configurations in Phase 1. In order to ensure the maximum efficiency in running the test cases after randomisation, Bayesian Optimisation (BO) as shown by Perdigão [Perdigão, 2024] will be used. Phase 2 will select the cases predicted to be most informative based on a selected acquisition function. BO causes estimates to be biased upwards as selection is no longer randomised. This will be corrected using IPW methods as shown by Hadad et al. [Hadad et al., 2021]. The independent variables of the experiment can be combined into 107 different treatment configurations and 1 control state, which is the current cloud configuration used. The same request will always be used, controlling for confounding variables that be introduced due to the size of the request or other characteristics. The outcome of processing this cloud request under these configurations will be measured in the time it took to execute the request. The outcome of the selected models are a conditional average treatment effect (CATE) where we compare the control state with the many different treatment states. The CATE can be represented mathematically as shown in formula 7. In order to perform this experiment we assume that the previous execution on the platform of the past configuration will not affect the next execution [Heckerman and Chickering, 1995] and that the influence of the independent variables on the dependent variable is non present unless explicitly stated [Heckerman and Chickering, 1995].

$$\text{CATE}(X) = \mathbb{E}[Y \mid X, W = 1] - \mathbb{E}[Y \mid X, W = 0] \tag{7}$$

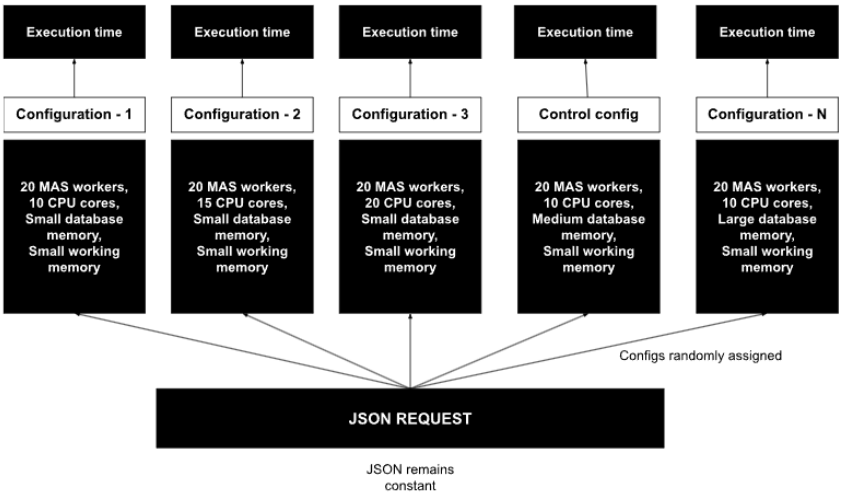

Figure 2: Causal DAG for Execution Time

To identify the causal affects of each of these independent variables on the execution time, since all variables in each configuration are expected to influence one another, we will be in need of a non-parametric model (equation 8).

$$\text{Time} = f(X_1, X_2, X_3, X_4, \ldots, X_n) \tag{8}$$

Based on past literature, the two causal learning models to use and compare are the Bayesian Neural Network and the Causal Forest. The aim is to estimate the conditional average treatment affect under each model so as to deduce the causal effects that each independent variable has on the dependent variable. Finally, a lift curve will be plotted to show the gained utility compared to the control configuration, capturing both time and money inside of utility. All of the analysis will be done using the python programming language with the use of the 'DoWhy' package which is specialised in experimentation. The methodology of this project actively uses the developments introduced by Kim and Pearl [Kim and Pearl, 1983] in regards to the use of DAGS and Bayesian networks.

4 Managerial relevance

The experiment intends to algorithmically identify the causal drivers of performance of their cloud systems, knowledge that is currently unknown. The Bank spends millions on computing power every year and therefore by understanding the causal drivers of high or low performance, The Banks' managers will be able to configure their credit models and processes to be as efficient as possible. Next, The Banks' current process for performance testing involves running the largest amount of test cases on the platform to identify its upper limit before it times out. This test expires and yields little insight. Understanding how these drivers influence performance will allow for timeless insights, used to estimate performance impacts. This process will prevent the need for continuous testing, improving efficiency.

On another level, thousands of firms use cloud computing [Li and Kumar, 2022] and they often do not understand the drivers of their computing resource use. Although this paper focuses on a specific use case, the methodology described can be easily replicated, switching the variables with that of another organisation, systemically uncovering the drivers of performance, optimising their computational load. This paper therefore is not restricted to the discussed use case, but is applicable to thousands of organisations across the world.

5 Academic relevance

The focus of the literature review is to ensure that the project fill a research gap, solidifying the fact that performance testing using causal learning has not been done previously, while learning from previous findings and methodology.

Examining foundational theories supporting the hypotheses, better cloud performance has been attributed towards increased CPU usage ([Guidi et al., 2020], system memory, increasing it by up to three times[Yang and Dulloor, 2019], database memory (and other techniques such as indexing) [Smith et al., 2021] and worker configurations (traffic distribution and load balancing among one or multiple workers) [Zeng et al., 2020]. Additionally, pure increases in these metrics do not scale performance linearly, but have been proven to be configuration and payload dependent [Wang et al., 2023].

Moving to existing research, Jayathilaka [Jayathilaka, 2017] developed an approach to attempt to locate the root cause of performance drops on cloud infrastructure. The authors used correlation analysis between the performance change and spikes in other metrics which is a different approach than what is proposed. Furthermore, Das [Das et al., 2022] applied causal learning models to identify performance drops in cyberphysical systems. While the methodology is similar, the context is different. One of the most relevant papers, authored by Hsu [Hsu et al., 2018], investigates finding the optimal cloud configuration set up, focusing on CPU, memory usage and other variables. The authors use Bayesian optimisation to locate efficient configurations. This paper has substantial overlap with the proposed topic. That said, the authors did not employ causal learning algorithms.

Next, examining previously used methodology. The paper by Athey and Imbens [Athey and Imbens, 2016] is one of the foundational papers on the Causal Forest. The causal forest partitions the features space much like a random forest model, but instead of splitting based on the RSS value, it splits based on the CATE estimates. Hundreds of papers have used the causal forest for similar projects such as Zhang [Zhang et al., 2024] who use causal forests to identify under performance in cloud computing nodes. The causal forest is ideal due to its 'honest' estimations. It does not reuse the same data to split the data as for CATE estimations, avoiding over fitting [Wager and Athey, 2018]. Finally, as shown in the systematic review of 133 papers by Rehill [Rehill, 2024] who focuses on the use of causal forests, the study found that causal forests are used in low dimensional datasets and rely on randomised controls, all of which are present in this study.

Bayesian Neural Networks (BNN) are another viable model as they account for complex relationships between nodes but instead of having a fixed weight in the neural network, the weight is on a distribution, representing the uncertainty [Titterington, 2004] such as the influence of variable of cloud configuration on execution time. Furthermore, the model does not impose a strict parametric form which reduces the risk of model mis-specification, which is crucial in causal inference [Xu et al., 2010]. Additionally, what was found in the paper by Chen [Chen et al., 2019] was that by employing the BNN, we can measure the dependency between variables, allowing for the dependent variable with high mutual on an independent variable to infer a stronger causal link, yielding interpretation around the relationship. Since the relationship of our variables is assumed to be nonlinear, the BNN is ideal to be used.

Having scoured the related academic publications, a gap has been identified. The community has never analysed performance changes using causal learning models. All of the aforementioned models have been proven to be well suited for the problem at hand and is the aim of this paper. This gap is valuable as it will give more accurate. interpretable and causal insights than previously conducted research.

6 Potential challenges

Operationally, given that the project is being undertaken from inside a large multi-national bank, there is the chance that timelines are delayed due to other priorities,

scheduling and data sharing. To overcome these issues, scheduling has taken place and a written contract has been signed to outline what can and cannot be shared.

The project is also faced with methodological challenges. Cloud system performance is driven by many aspects, not all of which are captured in this project. No other variables can be adjusted from the platform side as a user, nor can the request level data be captured either. Confounding remains a major issue in this causal paper and conditioning will be used but may not always be possible. Next, the methodology lacks an explicit treatment across all groups. Each configuration will have a change compared to the previous but there is no one treatment between treatment and control groups. This makes the experiment more complex and harder to interpret. Additionally, since we have one request and 108 configurations, it is unsure whether the sample size will be big enough to train a causal model and if so, how many repetitions are needed of each configuration. An insufficient sample size will result in high variance and poor power.

7 Timeline

Table 1 represents the required tasks for both the 'Bank' and the University. This is a summary of all technical and formal tasks required for the project. The majority of the tasks have already begun. All related stakeholders have been contacted and have the dates planned in their agendas to ensure a smooth process. As can be seen, the goal of the timeline is to have the complete project written and under draft review by the end of May, allowing for two weeks of feedback and adjustments prior to the final submission. At the end of the project, the goal is to implement the solution into The Bank' infrastructure. For any further questions, please inquire.

Tasks	**Deadline**
Contact all stakeholders for one month run worth of data for each model	07/02/2025
Literature review	09/02/2025
Start proposal RQ, PS, theoretical and business reasoning	09/02/2025
Ensure that all data is in JSON format for Liza	14/02/2025
Anonymize any test cases	14/02/2025
Scope of testing (which RS)	14/02/2025
Merge all data and clean / prepare the data	21/02/2025
Build randomisation script	21/02/2025
Select JSON test cases	21/02/2025
Send draft proposal questions to Jason	25/02/2025
Thesis proposal due	05/03/2025
Meet with Jason to discuss outcomes of proposal	05/03/2025
Execute test cases with performance measures recorded	10/03/2025
Put the data in GCP	10/03/2025
Train and tune	25/03/2025
Get sign-off from Anil for shared Git repo, LaTeX, etc.	02/04/2025
Meet with Jason to discuss outcome of model	15/04/2025
Methodology and results	15/04/2025
Implications, conclusions, and limitations	01/05/2025
Draft thesis review	25/05/2025
Thesis deadline	15/06/2025
Thesis defense	15/07/2025
Implement results into ops	15/07/2025

Table 1: Project Task List and Deadlines

References

[Athey and Imbens, 2016] Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

[Chen et al., 2019] Chen, J., Feng, J., Hu, J., and Sun, X. (2019). Mutual information in bayesian neural networks for cloud performance. *Journal of Cloud AI Systems*, 7(2):50–67.

[Das et al., 2022] Das, A. et al. (2022). Causal inference in cyber-physical systems: A cloud performance study. *Journal of Applied Cloud Computing*, 9(1):34–56.

[Ellis and Wong, 2008] Ellis, B. and Wong, W. H. (2008). Order space sampling for bayesian networks. *Journal of Machine Learning Research*, 9:355–380.

[Guidi et al., 2020] Guidi, G., Becchi, M., Barker, K., and Hoisie, A. (2020). 10 years later: Cloud computing is closing the performance gap. *arXiv preprint arXiv:2011.00656*.

[Hadad et al., 2021] Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15):e2014602118.

[Heckerman and Chickering, 1995] Heckerman, D. and Chickering, D. M. (1995). A bayesian approach to learning causal networks. *Machine Learning*, 20(3):197–243.

[Hsu et al., 2018] Hsu, C. et al. (2018). Optimizing cloud configurations: A bayesian approach. *Journal of Cloud Systems Engineering*, 10(4):98–120.

[Jayathilaka, 2017] Jayathilaka, A. (2017). Root cause analysis for cloud performance drops. *International Journal of Cloud Computing*, 15(3):123–140.

[Kim and Pearl, 1983] Kim, J. H. and Pearl, J. (1983). Title unknown. *Journal Unknown*.

[Li and Kumar, 2022] Li, B. and Kumar, S. (2022). Managing software-as-a-service: Pricing and operations. *Production and Operations Management*, 31(6):2588–2608.

[Makhlouf, 2019] Makhlouf, R. (2019). Cloudy transaction costs: A dive into cloud computing economics. *Journal of Cloud Computing: Advances, Systems and Applications*, 8(1):1–16.

[Perdigão, 2024] Perdigão, D. (2024). Bayesian optimization for performance testing in cloud systems. *Cloud Performance Journal*, 18(2):200–215.

[Qureshi et al., 2020] Qureshi, M. S., Qureshi, M. B., Fayaz, M., Zakarya, M., Aslam, S., and Shah, A. (2020). Time and cost efficient cloud resource allocation for real-time data-intensive smart systems. *Energies*, 13(21):5706.

[Rehill, 2024] Rehill, P. (2024). How do applied researchers use the causal forest? a methodological review. *Statistical Applications in Data Science*, 15(2):250–265.

[Smith et al., 2021] Smith, J., Doe, J., and Brown, M. (2021). Optimizing cloud computing performance with advanced dbms techniques: A comparative study. *Journal of Cloud Computing: Advances, Systems and Applications*, 10(4):123–145.

[Titterington, 2004] Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, 19(1):128–139.

[Wager and Athey, 2018] Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

[Wang et al., 2023] Wang, Q., Zhang, L., Xu, B., and Zhou, W. (2023). Container resource allocation versus performance of data-intensive applications on different cloud servers. *Future Generation Computer Systems*, 138:65–78.

[Xu et al., 2010] Xu, Y., Daniels, M. J., and Winterstein, A. G. (2010). Nonparametric bayesian models for high-dimensional data. *Bayesian Analysis*, 5(4):755–778.

[Yang and Dulloor, 2019] Yang, J. and Dulloor, S. R. (2019). Intel optane dc persistent memory: Challenges and opportunities. *Proceedings of the IEEE*, 107(12):2327–2342.

[Zeng et al., 2020] Zeng, R., Liu, Y., Wang, W., and Chang, X. (2020). Performance optimization for cloud computing systems in the microservice era: State-of-the-art and research opportunities. *IEEE Transactions on Cloud Computing*, 8(3):735–753.

[Zhang et al., 2024] Zhang, C., Yao, R., Qin, S., Li, Z., Agrawal, S., Mishra, B. R., Tran, T., Ma, M., Lin, Q., Chintalapati, M., and Zhang, D. (2024). Applying causal forests to cloud computing node performance. *Journal of Cloud Computing*, 12(1):100–115.