Visualization and Machine Learning for Data Center Management

Alina Chircu[1], Eldar Sultanow[2], David Baum[3], Christian Koch[4], and Matthias Seßler[5]

Abstract: In this paper, we present a novel tool for data center management that incorporates data visualization and machine learning capabilities. We developed the tool in the context of an action design research project conducted at a large government agency in Germany, which hosts three highly available data centers containing more than 10,000 servers. We derived the requirements for the tool from qualitative interviews with agency employees who are familiar with monitoring the data center infrastructure as well as from a review of existing data center and other large infrastructure monitoring solutions. We implemented a web-based 3D prototype for the tool as an Angular 6 application running on Node.js, and evaluated it with the same employees. Most participants preferred the new tool, which provided a significantly better option and enabled visualization of historical data for all server instances at the same time, as well as real-time charts. Planned improvements will take advantage of the full potential of machine learning for time series forecasting.

Keywords: Data Center, Forecasting, Machine Learning, Monitoring, Response Time, Time Series, Utilization, Visualization

1   Introduction

Data centers – physical facilities and the associated information technology (IT) infrastructure needed to host applications and data – are essential for an organization's success in the digital age [GRD13]. They are becoming even more important as organizations embrace digital transformation and embed digital technologies (such as web, cloud, or mobile technologies, Internet of Things, artificial intelligence, robotics, and the like) in all their business processes. Respondents to a 2019 survey indicate their organizations manage, on average, 12 data centers – a number expected to increase to 17 in three years and to include both cloud and on premise facilities [Kl19]. Whether they use an on-premises, a cloud-hosted, or a hybrid model, data centers now host increasing amounts of data as well as increasingly complex software and hardware such as "artificial intelligence engines, vast cloud ecosystems, blockchain solutions, advanced connectivity architectures, high-performance computing platforms, and more" [Kl19]. New solutions such as data analytics, machine learning, and artificial intelligence are among the emerg-

---

[1] Bentley University, 175 Forest Street, Waltham, MA 02452-4705, USA, achircu@bentley.edu
[2] Capgemini, Bahnhofstraße 11C, 90402 Nuremberg, Germany, eldar.sultanow@capgemini.com
[3] Business Informations Institute, University of Leipzig, Leipzig, Germany, david.baum@uni-leipzig.de
[4] Technische Hochschule Georg Simon Ohm, Nuremberg, Germany, christian.koch@th-nuernberg.de
[5] IT Dept. of Germany's Federal Employment Agency, matthias.sessler@arbeitsagentur.de

ing innovations [GRD13, Kl19] that can help ensure operational efficiency, service reliability, and security of data and applications in the increasingly complex data center environment.

In this paper, we present a novel tool for data center management that incorporates data visualization and machine learning capabilities. We developed the tool in the context of an action design research [Se11] project conducted at Germany's Federal Employment Agency (thereafter referred to as FEA). FEA manages a network of branch offices and local agencies in order to provide services to individuals and firms in a timely and efficient manner. Its employees use more than a hundred applications, accessed through 170,000 devices. Furthermore, millions of citizens use its online portal. The functionality of the provided software is implemented by a myriad of interconnected services, hosted in three highly available data centers comprising more than 10,000 servers altogether. The complexity of the environment and the high criticality of the services require extensive and efficient monitoring. This includes identification of abnormal behavior of any server at runtime (including temporal and recurring patterns) such as a failure of the system or a malfunction that will lead to a system failure. Thus, quick and early detection of abnormal behavior is critical. FEA's existing data center monitoring tool is designed for manual data analysis (drill-ups, drill-down, zooming into parts of data that seems to be abnormal based on peaks or dips in the data). In order to fix the issue it has to be classified (e.g. large response times, huge number of requests, connection pool, huge number of threads, huge number of failures, etc.) and the affected servers or clusters have to be localized in order to know which servers need maintenance. The new tool presented in this paper supports the user in detecting abnormal behavior, in classifying and localizing the issue, and in predicting future behavior.

This paper is organized as follows. Section 2 summarizes related work that is important for the development of our proposed tool. Section 3 briefly reviews the action design research methodology used in this research project. Section 4 presents the details of the new data center management tool – including gathering requirements from users, developing the tool, and evaluating it. Finally, Section 5 presents conclusions, limitations and future research directions.

2   Related work

As the importance of data centers for today's businesses has been increasing, researchers and practitioners alike started investigating data center management challenges and corresponding solutions [Mo05]. The research in this area recognizes the unique nature of data centers, which requires managers to consider both traditional IT problems such as performance monitoring, server management, and security, as well as facilities-related problems related to power consumption and temperature management [Mo05]. As a solution to data center complexity and fragmentation of operations, tools, and information, data center infrastructure management tools emerged as a way to better measure,

monitor, manage and control data center capacity, service delivery, resources, and assets [GRD13]. Analytics, machine learning and artificial intelligence are emerging as key solutions for data center operation management [GRD13, Kl19, LBB18, Mo05]. For example these methods can help better manage service delivery (analyzing and predicting downtime and performance, conducting root cause analysis, and designing resilience in case of outages) [GRD13] or power consumption and cooling costs given a specific workload [Mo05]. In particular, machine learning has been used to predict resource utilization, quality of service (response time), and workloads and to help automatic job and resource allocation [Be13].

The need for visualizing the load and status of a data center has been recognized and dealt with in a limited number of research papers so far. For example, Fisher et al. provide a solution called "Visual-I" for visualizing the status of networks and applications residing in a large data center [Fi08]. The system, developed based on requirements collected during a case study, allows "human users to rapidly assess the health of the system – quickly identifying problems that span across components and solving problems that challenge fully autonomic and machine-learning based management systems" [Fi08]. In addition, several US patents have been submitted for methods to visualize the health of data center objects [WB06] or to analyze capacity and load in a data center [BMS06].

Despite these developments, the visualization of large infrastructures such as data centers remains challenging. Established visualization, analytics and monitoring frameworks such as Elastic Stack (including Elasticsearch, Logstash, and Kibana), Grafana, and Graphite, as well as services or service platforms such as AppDynamics, Dynatrace, and New Relic rely on two-dimensional visualizations, which are not very useful due to the complexity of the data center infrastructure. A small number of papers focus on three-dimensional (3D) visualizations of infrastructure status and events. For example, Taylor, Brooks and McHugh present a 3D-capable visualization tool that shows "historical network flow data per port of an individual host machine or subnet on a network over time" [TBM08]. In a related domain, researchers have developed solutions for visualizing the structure of very large software systems – for example using the city metaphor [We10] and the recursive disk metaphor [MZ15]. Schilbach extends the recursive disk metaphor by using animation to visualize runtime behavior of software systems [Sc18]. Similar modifications are applied to the city metaphor to visualize performance data in real time [Og17]. However, a solution for three-dimensionally visualizing the load of servers that reside in a data center, grouped in clusters and by specialized function, does not yet exist. We describe the development of such a solution in this paper.

Another relevant area for developing visualization solutions is enterprise architecture (EA), which represents "the organizing logic for core business processes and IT infrastructure reflecting the standardization and integration of a company's operating model" [RWR06]. The EA concept is important because it focuses on models that include both the business logic and the technology details, as well as, in some cases, other organizational elements such as the organization structure and human actors and their roles, not

just the technology details. The study of EA has become more popular in recent years [BS16, BSC17, BSC18] as EA is becoming increasingly relevant in our todays agile world [BSC18]. EA is not only relevant for research, but also for practice [BSC17], since it helps create architectural templates, called reference architectures (RA), for various industry domains such as the aviation industry, the maritime industry, the telecommunications industry, the pharmaceutical, life science and healthcare industries, and others [BS16, BSC17, BSC18]. In particular, EA can help create business analytics and data science RAs [BS16, BSC18]. In this paper, we use the EA concepts in a concrete case and describe the resulting RA.

3   Methodology

This paper employs an action design research (ADR) methodology, which is appropriate for building and evaluating IT artifacts [Se11]. ADR is based on both action research [Ba97, BM04] and design science research [Se11, VKP04] as follows. Action research is a methodology used for diagnosing and solving real-world problems faced by organizations, implementing changes, evaluating them, and learning from the process [Ba97, BM04]. Action research relies on an iterative process that links theory and practice and supports collaboration between and learning by practitioners and researchers [Ba97, BM04, Se11]. Design science research is a methodology for designing, building and evaluating IT artifacts and developing design principles [Se11, VKP04]. ADR combines these two methodologies in a research process consisting of several phases: (1) problem formulation (based on theory and practice), (2) building, intervention and evaluation of the IT artifact in a specific organizational context (with the corresponding context influences and related interactions between researchers and practitioners), (3) reflection and learning (based on actions taken in previous phases), and (4) formalization of learning (through generalized outcomes). While the formalization of learning phase has to happen at the end of the research process, the first two phases of the ADR process can be repeated as needed in order to improve the IT artifact, and reflection and learning can happen after each one of the preceding stages as well [Se11].

In this paper, we report on phases 1-3 of the ADR process we conducted for developing a novel tool for monitoring and visualizing data center metrics. As our IT artifact development work is ongoing, we anticipate adding more functionality and conducting several more iterations of building, intervention and evaluation, followed by reflection and learning, before we can draw final generalized outcomes and conclude the process.

4   Developing the data center monitoring tool

4.1   Understanding the requirements

As described in the introduction, the infrastructure of the organization we study, FEA, consists of three data centers with over 10,000 servers. Each data center encompasses various clusters, which in turn contain many servers. Each server has several runtime metrics (including response time, number of requests, connection pool, number of threads, number of failures, etc.). To monitor these metrics, the data center operations department uses a manual analysis tool (see Figure 1) five to six times a week. If ad-hoc requests are made (such as "our application or service is too slow"), the operations team uses the tool more often. However, the existing tool does not provide the capability to monitor all this information at the same time.
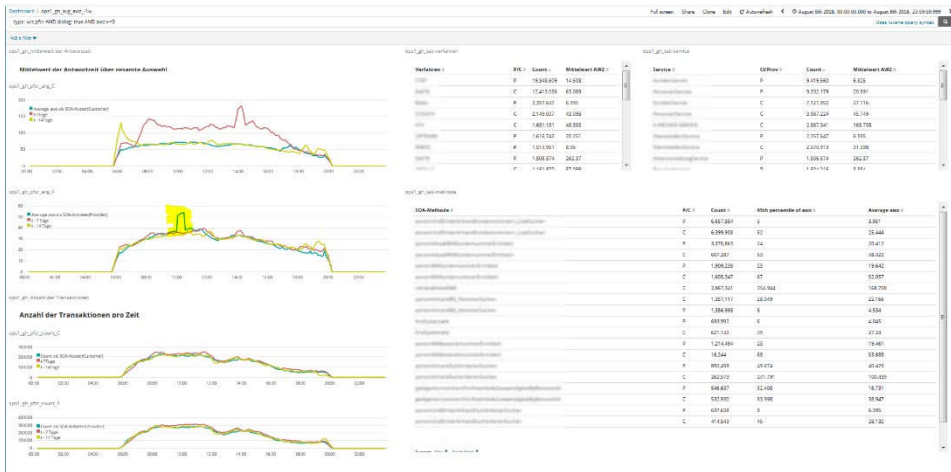

Figure 1. The tool currently used in production (showing the drill-down view for one server)

To get a better understanding of the requirements for an improved tool, we conducted qualitative interviews with seven employees of FEA who are familiar with monitoring the data center infrastructure. We asked them about the concrete tasks they have to per-form, the problems they have with the current tools, and the features they wish they had.

To understand each employee's goals and work processes in analyzing the data center metrics – which define some of the main requirements a monitoring tool should fulfill - we asked several questions. The respondents' answers to the question "What goals do you pursue / what do you have to deliver and/or to ensure?" included the following goals:

- *I have to recognize anomalies / faults in FEA's large infrastructure.*

- *I have to produce a more reliable and faster qualification of the operating state.*
- *I need to enable faster localization of the cause of an anomaly / disorder.*
- *I need to eliminate disturbances and to inform the relevant / affected it process-es (IT departments) about existing disturbances and about the causes.*
- *I need to increase the ability of predicting failures of individual instances.*

The respondents' answers to the question "How do you use the current tool to accomplish this?" included the following work processes:

- *I look for peaks, zoom in, and evaluate individual logs.*
- *I have to search for information in several tools (dashboard views) and I need to combine this information to locate the source of a reported problem.*
- *I search manually for contexts such as: Within a process or area, the response time of a service increases. Why is that? - In order to find out, I have to use the Kibana Dashboard to inspect the affected time period. I search for indicators that point to the cause of the problem. This might be for example another process, started by another department that shares the common resources.*
- *The ability to predict failures on individual instances is currently not possible with the tool.*

The question "What are the problems with the current tool?" generated the following statements:

- *The constant drill-up and drill-down is cumbersome – you do not have everything at a glance.*
- *Expert knowledge is required for each tool. Moreover data analysis know-how is required, but it is not available.*
- *This is a fairly time-consuming manual search process.*
- *Predictability is not possible. Even worse is that we (in the operations department) are often not even able to locate problems and their causes.*

In order to gain insight into possible future requirements, we asked the question "What other features would the future system need to fully round out the entire work in the company?" This generated the following list of desired features:

- *It would be great to have a mechanism in the application that draws attention to anomalies.*
- *A system that conveniently allows a human acquisition/cognition of such anom-alies, thanks to advanced visualization techniques. A feature that does not force you to search for a needle in a haystack by drilling up and drilling down.*
- *A machine learning based classification of service quality would be great.*
- *Incorporating machine learning for the localization of anomalies sounds like a promising feature to me.*
- *The ability to make "intelligent" suggestions such as "batch X of IT department Y might be the cause of the delays in IT department Z we are seeing" would be brilliant*

- *The future tool should not only apply machine learning for time series analysis and prediction. It should also predict and alert future expected outliers / peaks.*

4.2   Implementation

Based on the requirements obtained through interviews (see section 4.1) as well as on our review of existing monitoring solutions for large infrastructures (see section 2), we implemented a web-based 3D prototype as an Angular 6 application running on Node.js. The first version of the prototype only supports one failure class, i.e., the response time metric. Figure 2 presents a screen shot of the prototype. To visualize the utilization of the infrastructure we adopt the 3D city visualization metaphor used in other large systems applications (see section 2). We depict each server as a house in the city visualization and we use a runtime metric as the height of the building (i.e. response time in our first prototype). The houses are clustered according to the hierarchical structure of the clusters and data centers. Each cluster has its own color to make the distinction easily. The area of each building is constant in order to avoid any biases during analysis. Following user input, we also made several updates to the initial prototype (for example, the outage of a server instance is visualized by an empty square; in addition, in order to prevent the situation when a peak makes the visualization unusable because small bars would not be visible anymore, we introduce a logarithmic scale). Apart from the limitation to one failure class, the prototype supports all requirements initially obtained from the users.
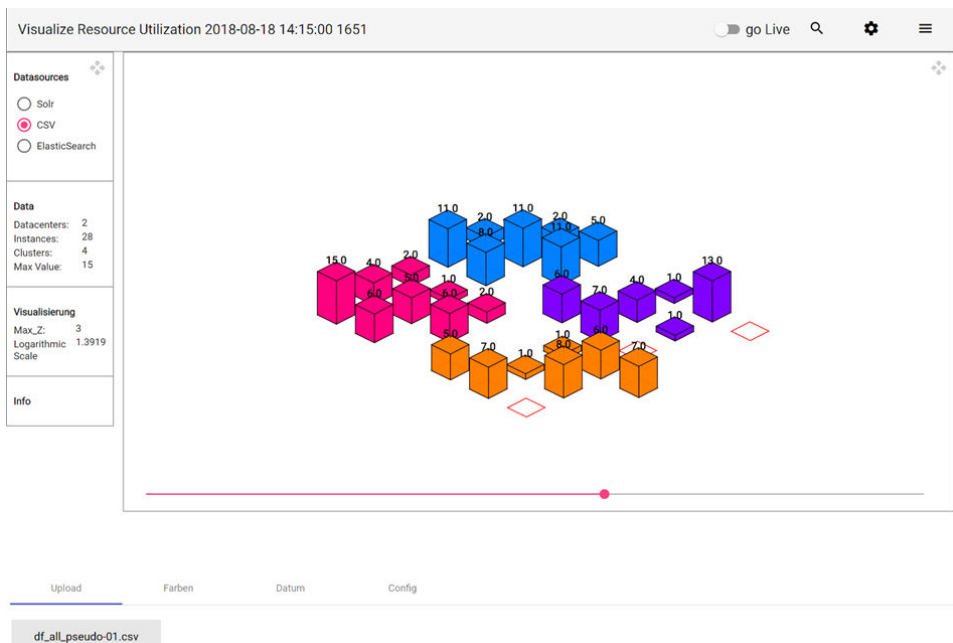
Figure 2. The prototype (2 data centers, 4 clusters, 28 server instances)

The prototype architecture is depicted in Figure 3. The architecture is organized into three layers: a bottom layer for data storage and retrieval, a middle layer for processing and analyzing the data, and a top layer for data visualization. Data transformation (in the bottom layer), also referred to as preprocessing, and data analysis (in the middle layer) are mainly performed using Python packages. The top data visualization layer including the graphical user interface uses JavaScript, since the frontend is an Angular 6 application. It also uses Three.JS and D3-3D, which are JavaScript-based frameworks that allow us to display the utilization data and the infrastructural status of the data centers, clusters and server instances in three dimensions. Our experience indicates that, for large data sets, the Three.js framework scales up extremely well as the amount of data to be displayed grows.
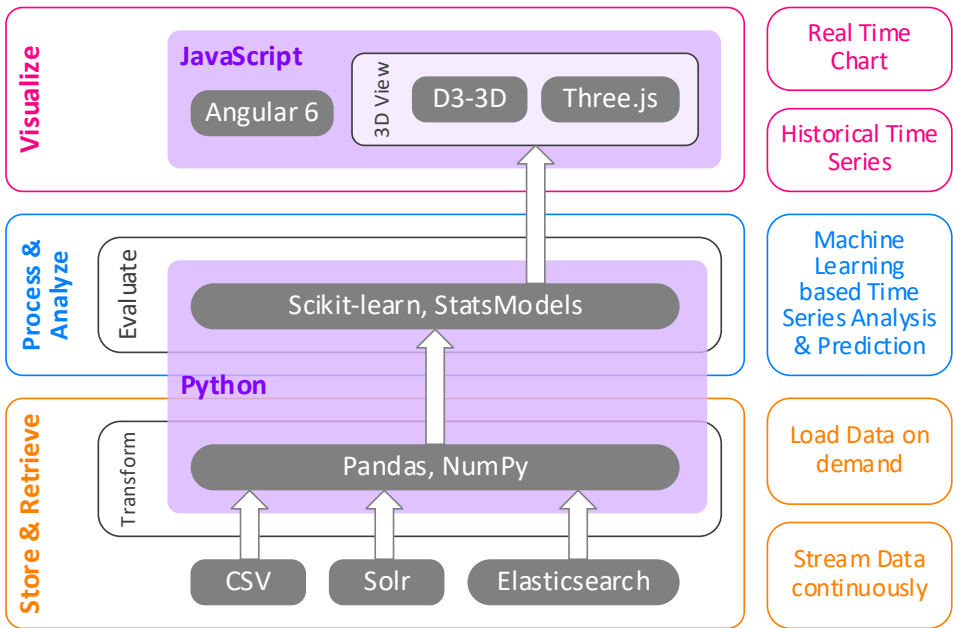

Figure 3: Architecture of the prototype

The prototype provides several options regarding data:

(1) The data is loaded from a CSV file where each row contains a timestamp, a value for the metric being visualized, and the IDs of the corresponding data center, cluster, and server instance. The user can analyze the utilization situation at different times by using a slider.

(2) The data, which is collected within the infrastructure (in our case it is collected from an enterprise service bus) is provided by the time series database Elasticsearch (Solr, Prometheus, or Graphite are other possible choices) and is continuously flowing into our system. The data is provisioned for visualization using the D3 Streaming JSON functionality, which enables us to show a Real-Time 3D Chart.

With a horizontal slider it is possible to switch between certain points in time or view the changes over time in fast motion. From our point of view this outperforms alternatives such as bar charts or conventional time-series representations for multiple reasons. First, our visualization scales up to many hundreds of server instances, which cannot be achieved with time-series representations for every single server. Second, the servers get clustered in a natural way representing the hierarchy of the infrastructure. Using the third dimension (time) helps the user recognize irregularities. Figure 4 presents an example of how this feature allows comparisons over time.
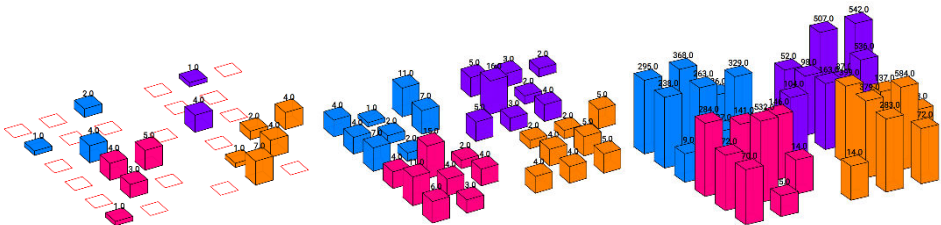

Figure 4: Visualizing changing load distribution (screenshots of the load distribution at three different points in time)

The prototype also provides several options for visualizing the data: a historical view, a real-time chart, and a dashboards with alerts and analyses based on machine learning processing, which enables time series forecasting and time series anomaly detection through supervised learning (this last feature is currently under development). We chose machine learning for our prototype because machine learning models have been success-fully applied to time series forecasting and are hypothesized to be better suited for real-world applications, especially for non-linear cases [BTL13, Ah10].

4.3   Evaluation

To investigate the usefulness of the presented visualization we conducted an ADR-based prototype evaluation with the same participants in the requirements interviews. We asked them to use the prototype to identify operational anomalies based on utilization / load data. The data was gathered from three data centers, four clusters and several dozen nodes from FEA's infrastructure. After completing the tasks with the prototype, the participants also used the existing tool that is currently used in production. We then asked them which visualization they prefer and why. Most participants preferred the new

system, and made the following comments regarding the improvements the prototype offers compared to the existing tool:

- *The prototype visualizes each instance in each cluster and data center at a glance. Here I get through faster.*
- *I have a visualization of all service-relevant metrics in one view.*
- *The three-dimensional view allows me to see more and to recognize more quickly how load behavior is related – and this without the need to correlate manually different views.*
- *The visualized time series data can be incorporated into a predictive analytics model. So we would utilize the load time series of those components that have failed, for predicting the failure of other components.*

Considering that the participants are much more used to the old tool, this is a very promising result. Nevertheless, there is still room for improvement. The participants made the following improvement suggestions during the ADR-based prototype evaluation (some of which we have already implemented):

- *Placeholders for non-present instances.*
- *Slider for adjusting the maximal height.*
- *Logarithmic scaling of the cuboids.*
- *Color configuration panel.*
- *Hover effects with details of the instance.*
- *Lazy loading the data.*
- *Real-time capability.*
- *Info and control panel for configuring the rules of parsing the CSV file (which columns are the relevant ones).*
- *Highlight important points on the x-axis.*
- *Interpolating values so that there are no extreme hops between values along the x-axis.*
- *A drop down menu for selecting IT departments.*

We have already implemented the first four points of the list. As soon as the remaining points have been implemented, we will re-evaluate the visualization.

5   Conclusions, limitations and future research

In this paper, we follow an action design research methodology to develop a novel tool that can help data center employees visualize complex infrastructures and monitor metrics of interest. We base our design on an analysis of the data center infrastructure of a large German federal agency, on interviews conducted with the agency's employees, and on a review of existing data center and other large infrastructure monitoring solutions. We present the architecture and implementation of the web-based 3D prototype for the tool, as well as the results of the prototype evaluation with real-world users – the majori-

ty of whom preferred the new tool, which provided a significantly better option and enabled visualization of historical data for all server instances at the same time, as well as real-time charts. To the best of our knowledge, an approach for three-dimensionally visualizing the load of servers that reside in a data center, grouped in clusters and by specialized function, such as depicted by our prototype, does not yet exist. Thus, the tool described in this paper adds to the emerging research on innovative solutions for data center management through visualization, analytics and machine learning.

Currently, the biggest limitation of our prototype is that it visualizes one metric. However, it is just a question of development effort to add more meta-data and allow classification of different performance issues. For instance, the meaning of the height of the buildings could be chosen at runtime by the user depending on what metric the user is interested in. In addition, we do not use the full potential of predictive analytics at this moment – it is still under construction. We are planning a full-fledged machine learning-based time series analysis and prediction functionality, which will be capable of forecasting the course of the time series and predicting outages, peaks, or other events that require attention. This component is not yet implemented, but considered as a next step for future releases.

References

[Ah10]   Ahmed, N. K.; Atiya, A. F.; Gayar, N. E.; El-Shishiny, H.:An Empirical Comparison of Machine Learning Models for Time Series Forecasting. Econometric Reviews, 29/5, 2010, pp. 594-622.

[Ba97]   Baskerville, R.: Distinguishing Action Research from Participative Case Studies. Journal of Systems and Information Technology, 1/1, 1997, pp. 24-43.

[Be13]   Berral, J. Ll.; Gavaldà, R.; Torres, J.: Empowering Automatic Data-Center Management with Machine Learning. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, 2013, pp. 170-172.

[BM04]   Baskerville, R.; Myers, M. D.: Special Issue on Action Research in Information Systems: Making IS Research Relevant to Practice: Foreword. MIS Quarterly, 28/3, 2004, pp. 329-335.

[BMS06]   Behera, S.; McBride, S. P.; Smith, J. D.: US Patent: Capacity and Load Analysis in a Datacenter, 2006.

[BS16]   Brockmann, C.; Sultanow, E.: Enterprise Architecture Management - Aktueller Stand und Ausblick. In (Mayr, H. C.; Pinzger, M., eds.): INFORMATIK 2016. Gesellschaft für Informatik, Bonn, p. 771, 2016..

[BSC17]   Brockmann, C.; Sultanow, E.; Czarnecki, C.: Unternehmensarchitektur in Forschung und Praxis. In (Eibl, M.; Gaedke, M., eds.): INFORMATIK 2017. Gesellschaft für Informatik, Bonn, p. 2041, 2017.

[BSC18]   Brockmann, C.; Sultanow, E.; Czarnecki, C.: Enterprise Architectures between Agility and Traditional Methodologies. In (Czarnecki, C.; Brockmann, C.; Sultanow, E.;

Koschmider, A.; Selzer, A., eds.): Workshops der INFORMATIK 2018 - Architekturen, Prozesse, Sicherheit und Nachhaltigkeit, Köllen Druck+Verlag GmbH, Bonn, p. 15, 2018.

[BTL13]   Bontempi, G.; Taieb, S. B.; Le Borgne, Y. A.: Machine Learning Strategies for Time Series Forecasting. In (Aufaure, M.-A.; Zim'anyi, E., eds.): eBISS 2012, LNBIP 138, Springer, Berlin, pp. 62–77, 2013.

[Fi08]   Fisher, D.; Maltz, D. A.; Greenberg, A.; Wang, X.; Warncke, H.; Robertson, G.; Czerwinski, M.: Using Visualization to Support Network and Application Management in a Data Center. 2008 IEEE Internet Network Management Workshop, 2008.

[GRD13]   Gilbert, P.; Ramakrishnan, K.; Diersen, R.: From Data Center Metrics to Data Center Analytics: How to Unlock the Full Business Value of DCIM. CA Technologies White Paper, 2013.

[Kl19]   Kleyman, B: How the Data Center Industry Is Supporting Digital Transformation in 2019.   Data   Center   Knowledge,   2019, https://www.datacenterknowledge.com/afcom/how-data-center-industry-supporting-digital-transformation-2019, accessed: 17/04/2019.

[LBB18]   Libri, A; Bartolini, A; Benini, L.: DiG: Enabling Out-of-Band Scalable High-Resolution Monitoring for Data-Center Analytics, Automation and Control. The 2nd International Industry/University Workshop on Data-center Automation, Analytics, and Control, Dallas, Texas, 2018.

[Mo05]   Moore, J.; Chase, J.; Farkas, K.; Ranganathan, P.: Data Center Workload Monitoring, Analysis, and Emulation. Eighth Workshop on Computer Architecture Evaluation using Commercial Workloads, 2005, pp. 1-8.

[MZ15]   Müller, R.; Zeckzer, D.: The Recursive Disk Metaphor – A Glyph-based Approach for Software Visualization. Proceedings of the 6th International Conference on Information Visualization Theory and Applications, SciTePress, 2015, pp.171–176.

[Og17]   Ogami, K.; Kula, R. G.; Hata, H.; Ishio, T.; Matsumoto, K.: Using High-Rising Cities to Visualize Performance in Real-Time. IEEE Vissoft 2017.

[RWR06]   Ross, J. R.; Weill, P.; Robertson, D.: Enterprise Architecture as Strategy: Creating a Foundation for Business Execution. Harvard Business Press, Boston, 2006.

[Sc18]   Schilbach, J.: Analyse, Erzeugung und Evaluation animierter Softwarevisualisierungen. Doctoral Dissertation, Leipzig University, Leipzig, Germany, 2018.

[Se11]   Sein, M. K.; Henfridsson, O.; Purao, S.; Rossi, M.; Lindgren, R.: Action Design Research. MIS Quarterly, 35/1, 2011, pp. 37-56.

[TBM08]   Taylor, T.; Brooks, S.; Mchugh, J.: Netbytes Viewer: An Entity-Based Netflow Visualization Utility For Identifying Intrusive Behavior. Mathematics and Visualization, 2008, pp. 101–114.

[VKP]   Vaishnavi, V.; Kuechler, W.; Petter, S. (eds.): Design Science Research in Information Systems, DESRIST.org, 2004 (created in 2004 and updated until 2015 by Vaishnavi,

V. and Kuechler, W.; last updated by Vaishnavi, V. and Petter, S. on December 20, 2017)

[WB06]   Wong, K. N.; Bednar, M.: US Patent: Method and Apparatus for Visualizing the Health of Datacenter Objects, 2006.

[Wet10]   Wettel, R.: Software Systems as Cities. Doctoral Dissertation, Università della Svizzera italiana, Lugano, Switzerland, 2010.