# SPEC Research — Introducing the Predictive Data Analytics Working Group

Poster Paper

André Bauer
University of Würzburg
Germany

Mark Leznik
Ulm University
Germany

Md Shahriar Iqbal
University of South Carolina
USA

Daniel Seybold
Ulm University
Germany

Igor Trubin
Capital One
USA

Benjamin Erb
Ulm University
Germany

Jörg Domaschka
Ulm University
Germany

Pooyan Jamshidi
University of South Carolina
USA

## ABSTRACT

The research field of data analytics has grown significantly with the increase of gathered and available data. Accordingly, a large number of tools, metrics, and best practices have been proposed to make sense of this vast amount of data. To this end, benchmarking and standardization are needed to understand the proposed approaches better and continuously improve them. For this purpose, numerous associations and committees exist. One of them is SPEC (Standard Performance Evaluation Corporation), a non-profit corporation for the standardization and benchmarking of performance and energy evaluations. This paper gives an overview of the recently established SPEC RG Predictive Data Analytics Working Group. The mission of this group is to foster interaction between industry and academia by contributing research to the standardization and benchmarking of various aspects of data analytics.

## CCS CONCEPTS

• **General and reference → Measurement**; **Metrics**; • **Computing methodologies → Machine learning**; **Artificial intelligence**; • **Information systems → Data management systems**; **Information storage systems**.

## KEYWORDS

SPEC, data analytics, data management, standardization, metrics, measurements, benchmarking

**ACM Reference Format:**
André Bauer, Mark Leznik, Md Shahriar Iqbal, Daniel Seybold, Igor Trubin, Benjamin Erb, Jörg Domaschka, and Pooyan Jamshidi. 2022. SPEC Research — Introducing the Predictive Data Analytics Working Group: Poster Paper. In *Companion of the 2022 ACM/SPEC International Conference on Performance Engineering (ICPE '22 Companion), April 9–13, 2022, Bejing, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3491204.3527495


*ICPE '22 Companion, April 9–13, 2022, Bejing, China*

ACM ISBN 978-1-4503-9159-7/22/04.
https://doi.org/10.1145/3491204.3527495

## 1 MOTIVATION

"Information is the oil of the 21st century, and analytics is the combustion engine". When Peter Sondergaard (Gartner Research) stated this phrase, the total amount of data created, captured, and consumed worldwide was 5 Zettabytes. Today, ten years later, that number has grown to 64.2 Zettabytes and is estimated to reach 181 Zettabytes by 2025[1].

Consequently, the research field of data based analytics—machine and deep learning and by extension, time series forecasting—has grown significantly in recent years as a means to make sense of the vast amount of available data. It has permeated every aspect of computer science and engineering and is heavily involved in business decision-making. For example, in the field of performance engineering, performance prediction is an instrument for controlling and improving the behavior of a system. Streamlining data analytics processes (DataOps) broadens and eases the access to data and applied models. Further, the striving for reproducible and reliable evaluations discourages "data analytics by the own workstation under the desk" approaches and demands generic architectures and software stacks.

The definition of such stacks poses a multitude of questions related to software and performance engineering: (i) The choice of low-levels of the infrastructure, including the storage medium, redundancy mechanisms, and file systems; (ii) The choice of a storage system suited for the type of queries issued by analytics/machine learning tools; (iii) The choice of specific mechanisms and procedures to be used for specific types of data, and the choice of the right tools for specific mechanisms; (iv) The choice of the suitable methodology for a specific problem.

To face these challenges, a multitude of solutions have been proposed. As there is no one-size-fits-all approach, these have to be analysed and evaluated. For this purpose, benchmarking and standardization are two powerful tools. On the one hand, benchmarking allows us to understand the solution of interest better and

[1] https://www.statista.com/statistics/871513/worldwide-data-created/

improve it continuously. On the other hand, standardization leads to innovation and the dissemination of knowledge, as it provides structured methods and reliable data.

For standardization and benchmarking, there are several associations and committees. One of them is SPEC (Standard Performance Evaluation Corporation), a non-profit corporation for the standardization and benchmarking of performance and energy evaluations. In this paper, we give an overview of the recently established SPEC RG Predictive Data Analytics Working Group[2]. The mission of this group is to foster interaction between industry and academia by contributing research to the standardization and benchmarking of various aspects of data analytics.

## 2 SPEC - STANDARD PERFORMANCE EVALUATION CORPORATION

SPEC is a nonprofit corporation for the standardization and benchmarking of performance and energy evaluations. It was formed at the initiative and under the Electronic Engineering Times (E.E. Times) sponsorship and through the joint development efforts of Hewlett-Packard Corp., Apollo Computer Inc., Sun Microsystems Inc., and MIPS Computer Systems Inc. The mission of SPEC is to provide the marketplace with a fair and useful set of metrics to differentiate the latest generation computing platforms. Since its inception in 1984, SPEC has become one of the most successful performance standardization bodies. The SPEC community has developed more than 30 industry-standard benchmarks for evaluating system performance across various application domains and has deployed thousands of benchmark licenses worldwide. SPEC publishes several hundred different performance results each quarter, covering a wide range of system performance disciplines. [1]

SPEC has established several board committees to address specific tasks and supports several benchmark development groups under the SPEC umbrella: the Open Systems Group (OSG), the High Performance Group (HPG), the Graphics and Workstation Performance Group (GPWG), the Research Group (RG), and the International Standards Group (ISG).

SPEC membership is open to all interested companies and currently includes computer hardware and software companies businesses, educational institutions, and government agencies.

## 3 RG PREDICTIVE DATA ANALYTICS WORKING GROUP

To bridge the missing links between the facets involved in data analytics, namely big data storage and provisioning, data versioning, and performance evaluation, the SPEC predictive data analytics working group was established in June 2021. The group's ambition is to standardize and benchmark the entire data lifecycle, i.e., the analytics/prediction methods and especially pipelines for data analytics ranging from big data storage and preprocessing to analytics and assessment, as well as to provide heuristics for the selection of tools, patterns, and infrastructure. Furthermore, the group members are interested in promoting the interaction between industry and academia by contributing research towards standardization and benchmarking of the different aspects of data analytics. For this, the group members investigate data analytics-related methodologies, systems, and metrics. Another important goal is to support open data and promote the reproducibility of experiments and benchmarking of data analytics methods. The interests of the group lie in but are not limited to:

(1) Performance modeling, analysis, testing, and prediction
(2) Performance analysis of ML Systems
(3) Resource autoscaling and reconfigurable systems
(4) Performance behavior in resource-constrained environments
(5) Changepoint and anomaly detection
(6) Time series analysis and forecasting
(7) Streamlining the data science process (DataOps)
(8) Benchmarking of big data infrastructure

In the following, we highlight some of the research areas.

*Performance Prediction.* Machine/deep learning approaches are utilized for performance prediction based on collected multi-channel performance data.

*Change Point and Anomaly Detection.* Statistical methods and machine/deep learning approaches are applied for detecting change points as well as anomalies/outliers. An example of our work can be found in our recently published paper [2] regarding the automated detection of change points in performance regression tests.

*Time Series Analysis and Forecasting.* Typically, collected performance data is available in the form of measurements over time. To this end, the group is interested in the analysis and forecasting/prediction of time series, where the following topics are of interest: *Feature Engineering*, *Clustering/Segmentation*, *Imputation*, *Synthesis*, and *Data Quality Assessment*.

*Streamlining the data science process (DataOps).* The data must be standardized and versioned to have a pipeline with interchangeable parts for all data analytics applications. Moreover, the deployment of such pipelines should be automated and standardized.

## 4 CONCLUSION

With this paper, we want to give an overview of the recently established SPEC RG Predictive Data Analytics Working Group, which aims is to investigate, understand, standardise, and benchmark the data analytics task.

## ACKNOWLEDGEMENTS

The authors want to acknowledge all the participants from member organizations and the SPEC HQ who have contributed to the overall success of SPEC. The name SPEC, together with its tool, benchmark, and service names, are registered trademarks of the Standard Performance Evaluation Corporation (SPEC).

## REFERENCES

[1] Walter Bays and Klaus-Dieter Lange. 2012. Spec: driving better benchmarks. In *Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering*. 249–250.

[2] Mark Leznik, Md Shahriar Iqbal, Igor Trubin, Arne Lochner, Pooyan Jamshidi, and Andrè Bauer. 2022. Change Point Detection for MongoDB Time Series Performance Regression. In *Proceedings of the 11th ACM/SPEC International Conference on Performance Engineering*.

[2]SPEC RG Predictive Data Analytics Working Group: https://research.spec.org/working-groups/rg-predictive-data-analytics/